

Towards a clean collider

The high-energy physics community has grand plans to probe deeper into the structure of matter and space-time. The proposal for a multinational linear collider merits strong support.

Although the Large Hadron Collider (LHC) is still years away from completion at CERN in Geneva, the particle-physics community is already mobilizing to push through the funding for its next project — a US\$6-billion linear collider to study collisions between electrons and anti-electrons. The announcement by the US Department of Energy (DOE) on Monday of its funding priorities for scientific facilities over the next 20 years is grist to their mill: the proposed Linear Collider ranks 13th out of 28 projects, in a list headed by the fusion experiment ITER (see page 108). The Linear Collider is classed first among the ‘mid-term’ priorities — mid-term being an indication that construction of this facility is still some way off.

But plans are progressing. In Paris next week, a steering committee will meet to establish a panel of ‘wise men’ to oversee the choice of accelerator technology for what is usually known as the International Linear Collider (ILC). Throughout 2004, the panel will review four competing accelerator designs, all at varying stages of development in laboratories around the world. By the end of the year, they will make their choice, and an international design project will be founded.

The particle-physics community is well-practised at such large-scale cooperation. CERN, established by nine European countries and now counting 20 European states as full members, celebrates its 50th birthday next year, marking a remarkable period of collaborative effort that has made possible research that no one country could itself afford to pursue. By financial necessity, the LHC project has expanded to include ‘guest’ states, including the United States, India and China. So the step to a formal international project is small, and the will is great. Speaking at a conference this summer, Peter Rosen of the DOE stressed that the ILC project must be an international effort “from inception”. A web-based petition in support of the ILC has been signed by more than 1,800 particle physicists worldwide.

The standard model of particle physics has been explored and verified more stringently than perhaps any other model in science. Occasional hints of something unexplained, something beyond the standard model, have evaporated. But the standard model is not the whole story — there are still issues to be resolved, such as the existence of the Higgs boson. The search for this elusive particle has been promoted as a major justification for building the LHC and for a substantial upgrade of Fermilab’s Tevatron accelerator.

Another dimension

What more can the ILC offer? The new machine, planned to stretch over 30 kilometres, will give physicists the chance to look beyond the discovery of the Higgs boson. This particle is likely to appear in the high-energy proton–proton collisions in the LHC. But that environment is ‘messy’: to pick out in fine detail the true nature of this particle (or particles, as it may be one of many), the precise, clean physics of the ILC’s electron–anti-electron collisions will be necessary. The interplay between the LHC and the ILC is a central plank in the physics case. The ILC won’t just be tidying up after the LHC, but, if built quickly enough, it will feed information back into refined analyses at the LHC.

There is likely to be more to investigate at the energy scales accessible to the ILC. For example, there may be another family of particles that mirrors the fundamental particles of the standard model. These are ‘supersymmetric’ particles, the lightest of which is a prime candidate for the dark matter in the Universe. As with the Higgs boson, the essential nature of this postulated particle should become clear at the ILC. There may also be extra dimensions of space, beyond the three of familiar experience: data from the ILC will tell us how many and how big they are.

The ILC will increase the energy of its collisions in stages. This makes sense from a design perspective, but will also bring a degree of flexibility to this project that has been impossible elsewhere. The ILC should be able to ‘home in’ on a particular collision energy, with the result, say, that copious amounts of Higgs bosons could be generated for study, or that a certain region of phase space in which the LHC has thrown up some anomaly can be investigated further. The machine could also be adapted at a later stage so that, instead of electrons and their anti-particles, the colliding particles are photons. Again this is another window on Higgs physics, and more besides.

Strong record

The United States will be expected to be a major player in the ILC project, and may even host the machine itself. Despite the DOE’s endorsement — and putting the issue of visas for visiting scientists aside — the situation in the United States is still complicated by the spectre of the Superconducting Supercollider, a US\$3-billion enterprise whose demise in 1993 shook the confidence of US particle physicists. Over the summer, that confidence was undermined once more, as reports from Fermilab suggested that its Tevatron accelerator may not deliver what was promised.

Meanwhile, in Europe, German physicists are still reeling from the halting of their bid to host a linear collider, which was a victim of a poor economic climate in that country. And the LHC project at CERN has suffered its own well-publicized problems, which resulted in a full external review last year. But those lessons have been learned well: far from being off-putting, that experience bodes well for the future. This physics community has, overall, a strong record in the management and delivery of accelerator-based projects, and its level of organization should inspire confidence in its ability now to deliver the ILC. Nevertheless, CERN may well be eyeing the ILC with caution: the laboratory’s LHC debts will not be paid off until 2010, and its management will be wary of any project that might lure away additional funds.

Physicists in the United States must now ensure that the ILC keeps its place on that DOE plan — which comes with the caveat that it will be reassessed periodically and its priorities may change. Worldwide, physicists must convince their politicians that the grand scheme of the Large Hadron Collider, followed by an International Linear Collider, offers a vital and vibrant programme of fundamental physics. Perhaps the possibility of spin-off technology from the development of the linear collider will tip the balance. With the estimated \$6-billion cost of the ILC spread over several years and shared by many nations, this is a bargain too good to miss. ■





Physics build-up
Energy department
sets out priorities for
US facilities
p108



Final journey
UK curators fear legal
move to repatriate
bone collections
p109



Error message
Dispute over potential
'breakthrough' in
hard-disk capacity
p110



Primate protest
Researcher quits as
Gibraltar macaque
cull wrecks study
p111

Consumers warned that time is not yet ripe for nutrition profiling

Erika Check

One day, information about your genome may well help you decide what breakfast cereal to eat. But that day's a long way off, the second International Nutrigenomics Conference in Amsterdam was told last week. In the meantime, researchers at the meeting heard, the emerging field badly needs a regulatory framework that will stop its first customers from being scared off.

Nutrigenomics researchers aim to learn how nutrients interact with genes to lead to health or disease. But people eat wildly different levels of nutrients over their lifetimes, and teasing apart the precise interactions is notoriously difficult.

The researchers who gathered in Amsterdam on 6–7 November were in optimistic mood, however. Their science is progressing quickly, and food industry executives have expressed interest in the idea of using genetic information to customize their products.

In January, the US National Institutes of Health used a 5-year, \$6.5-million grant to create a National Center of Excellence for Nutritional Genomics at the University of California, Davis, and the Children's Hospital Oakland Research Institute (CHORI) in Oakland. In July, the European Commission set up the European NutriGenomics Organisation to coordinate work. Now the Netherlands looks set to embark on a \$20-million nutrigenomics project, jointly funded by the government and the food industry.

But some researchers warn that the field is in danger of developing too quickly. They want experts to back off from the sometimes-extravagant claims for the field's potential, and instead to sit down and patiently work out a scientific vision and ethical framework for the discipline.

"Our aim is to bring the field a little bit back down to Earth, because people tend to start with a lot of science fiction," says Michael Müller, a genomicist at Wageningen University in the Netherlands who helped to organize the meeting.

The main fruits of this field are still years away, researchers say. So far, most of the studies on profiling gene expression — measuring genome-wide responses to nutrients —



Looks good, tastes good, and one day individuals may know exactly how much good it does them.

have been done in mice. And much more work is needed on the basic mechanisms by which nutrients turn genes on or off. But that hasn't stopped a handful of companies from selling nutritional profiles directly to consumers over the Internet.

The companies test a tissue sample — such as a cheek swab — from a "patient". The patient can choose which genetic profile he or she wants to learn about, for example skin ageing or susceptibility to osteoporosis. The company then gives the patient a "personalized profile" based on its tests for single nucleotide polymorphisms (SNPs): genetic variants that have been linked to disease. For instance, one company, GeneLink of Margate, New Jersey, tells people what vitamins they should take, based on SNPs involved in cellular responses to certain toxins. GeneLink declined to comment on its products.

But many scientists argue that it's far too early for most of these tests to be useful. "The idea of marketing any individual genetic test at this point assumes there is information to justify the use of that test, and we really don't have evidence that any single genetic marker

carries enough information to guide dietary treatments," says Ronald Krauss, director of atherosclerosis research at CHORI.

The direct-to-consumer tests also raise ethical issues that affect the whole field. For instance, some companies sell the results of their genetic profiles to other firms, which use the information for research on genes and disease. Although consumers must give their consent, they may not necessarily understand what they're agreeing to, says ethicist David Castle of the University of Guelph. Castle is collaborating with the University of Toronto Joint Center for Bioethics in soliciting comments on a joint working paper on ethics and nutrigenomics.

At the nutrigenomics meeting, Castle argued that even though the field is very young, scientists must begin talking to the public about such issues.

"This technology could end up affecting something that every person does every day, which is eat," Castle says. "It's not a situation where you want to roll out the science and the products and then go back and ask people how they feel about it."

Antarctic research frozen out of Japan's budget plans

Keiko Kandachi, Tokyo

Japan's finance ministry is set to pull the plug on a ¥40-billion (US\$360-million) project to build a research ship for Antarctic exploration.

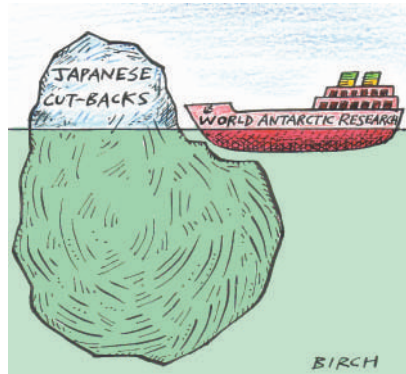
Ministry officials say that the plan lacks the popular appeal to justify its cost. But Earth scientists contend that the ship is essential to Japan's Antarctic research programme. On 7 November, researchers were joined by a range of celebrities at a rally in Tokyo in support of the project.

The country's current ice-breaking boat, the 21-year-old *Shirase*, has only about four years of working life left, researchers argue. But the finance ministry is reluctant to fund a replacement — despite pleas from the Council for Science and Technology Policy, Japan's highest science-policy body, which restated its support for the ship on 17 October.

Finance-ministry officials say that next year's budget, to be announced next month, is unlikely to include initial funding for the ship. "We don't think the Antarctic programme has the nation's understanding and support to justify the large budget," says Hideaki Imamura, deputy budget director at the finance ministry.

Other projects, such as the ¥60-billion Earth-drilling ship, *Chikyu*, which is currently under construction, seem to have overstretched the resources available for ocean research in Japan.

Loss of the ice-breaker would impair international research in Antarctica, says Chris Rapley, director of the British Antarctic Survey. Even a delay would damage programmes that depend on continuously collected data, says Okitsugu Watanabe, director-general of the National Institute of Polar Research in Tokyo. But he concedes that Japanese Antarctic research could continue without the new ship. "We could negotiate to use ships from other countries," he says. ■



Future view: Spencer Abraham reveals a list of the 28 facilities that US researchers need the most.

US reveals physics wish-list in bid for scientific frontier

Geoff Brumfiel, Washington

It must have felt as though Christmas had come early for US physicists this week, when the Department of Energy announced a list of 28 scientific facilities that it wants to build in the next 20 years. But it remains to be seen if the department's Office of Science can raise the billions of dollars needed to complete the projects.

At the top of the department's near-term priorities is the international fusion project, ITER, which the United States rejoined this year. Next is a proposal for a major scientific computing facility, to house supercomputers more powerful than anything currently available for non-military research. Then — equally ranked — come a rare-isotope accelerator, a satellite to probe the Universe's dark energy, a high-power X-ray light source, and a facility that would produce and tag proteins for research worldwide.

"This is our plan to keep the United States at the scientific frontier," says Spencer Abraham, the energy secretary, who announced the wish-list at a 10 November lunch at the National Press Club in Washington DC.

The list — the brainchild of Ray Orbach, head of the Office of Science (see *Nature* 424, 357–358; 2003) — is the first of its kind in almost 20 years. A somewhat similar plan was unveiled in the 1980s by Alvin Trivelpiece, then head of the Office of Energy Research. Trivelpiece welcomes the new plan. "It's a well-organized, thought-through proposal," he says.

The plan ranks facilities and upgrades

from across the department's laboratories into near, mid- and far-term projects. A US\$6-billion linear collider, to explore particle physics by colliding electrons and positrons, is the top mid-term priority, slated to begin around 2010. The highest-ranked far-term priorities include a US\$400-million upgrade to the National Synchrotron Light Source at Brookhaven National Laboratory in New York state and a neutrino beam that would be ten times more intense than anything currently available.

Orbach's office ranked facilities on the list by their scientific merit and their relevance to the energy department's mission, officials say. The result closely mirrors the suggestions of the agency's scientific advisory groups and laboratories, says Martha Krebs, president of the consulting firm Science Strategies in Los Angeles and head of the science office under President Bill Clinton.

Abraham says that he hopes the list will help raise the profile of the agency's science mission among the public and in Congress. "I didn't know the Office of Science as well as I do today," he says, "and I don't think a lot of members of Congress do either."

But physics lobbyists are already worried that few of the plan's projects will come to fruition. "When you have a nation running a deficit of half-a-trillion dollars, it's going to be hard to get anything to increase," says Michael Lubell, a physicist at the City University of New York, and head of public affairs at the American Physical Society. ■

www.er.doe.gov

Bone archives face prospect of dispersal

Jim Giles, Cambridge

In the days when the sun never set on the Empire, British ethnographers and anatomists freely collected skeletal remains from around the globe. But for years now, indigenous communities — especially those in Australia — have been asking for the remains of their ancestors back. And last week, a panel of academics and curators advised the British government to make legal changes that would open the way to such repatriations.

The panel, chaired by Norman Palmer, an expert in commercial law at University College London, said in its 5 November report that any remains should only be held with the consent of their direct descendants or leaders of a group of their descendants.

About half of the contents of Britain's two main collections of human remains — at the University of Cambridge and the Natural History Museum in London — originate from outside the country and would be subject to any new repatriation laws.

Aboriginal remains from Australia are likely to be among the first to be claimed back from Britain if the law is changed. Indigenous Australians argue that the remains were stolen, and want to lay them to rest in accordance with their cultural traditions. Specimens from Australian museums have already been returned.

The UK report is “a step towards true justice”, says Rodney Dillon of the Aboriginal and Torres Strait Islander Commission, an Australian group that campaigns for indigenous peoples' rights. “These remains were stolen. The remains are of human beings and they deserve to be buried like everyone else,” Dillon says.

Researchers say they understand the campaigners' feelings, but that the break-up of the collections would undermine efforts to better understand human evolution and the history



Heading home: skulls like these from the Duckworth collections may soon be joining their ancestors.

of the indigenous peoples themselves.

Marta Mirazón Lahr, an expert in human evolution and director of the Cambridge-based Duckworth collections, has used the bones in the collections to investigate how human appearance has changed over the past 100,000 years. The collections house the remains of more than 16,000 people, kept in rows of brown boxes. Neat black handwriting in yellowing books records the origin of each bone. “Skull of Aboriginal king, found in hollow of gum tree,” reads a typical entry.

“These remains give us so much information about the diversity of people,” says Lahr, who is dreading the implementation of the panel's findings. “This would be more than losing a few skulls. It would be like burning a library.”

The government has pledged to respond

to the report in January, and says it will consult with museums and universities before drafting any new legislation. But Tony Blair, the British prime minister, agreed in 2000 to cooperate with Australia over the repatriation of remains, and the political momentum appears to favour change.

In the United States, the 1990 Native American Graves Protection and Repatriation Act has led to fierce legal battles between indigenous groups and scientists over the use of human remains (see *Nature* 419, 5; 2002). It is unclear how many indigenous peoples from around the world would be interested in seeking the repatriation of remains from Britain's collections. Lahr says that most African objects in her collection are from Egypt, and she does not expect groups there to seek their repatriation. ■

African states pledge increased spending on research

Michael Cherry, Cape Town

Thirty-four African countries have committed themselves to spending at least 1% of their economic output on science and technology within five years.

The pledge was made in Johannesburg last week at the inaugural meeting of the ministerial council on science and technology, part of the New Partnership for Africa's Development (NEPAD) initiative.

Currently, Egypt is the only African country to spend 1% of its gross domestic product on research and development, although Algeria is close behind. South Africa, which will chair the ministerial

council responsible for implementing the NEPAD science and technology programme, estimates its current spending at 0.7%, and hopes this will rise to 1% within four years.

Wieland Gevers, president of the Academy of Science of South Africa, welcomed the announcement as “a huge step forward”, but added that success “would depend on how purposefully funds were used”.

Gevers hopes that the commitment will help attract aid from rich countries to support research in Africa — although he warned that Africans should not be junior partners in any resulting work. He said that an alliance of 10 African science academies

will meet next month to consider how they can move towards the goal.

Biotechnology, science and technology for manufacturing, and energy and water research were among the 12 fields emphasized at the meeting. The aim is to create a centre of excellence in each field.

The African Laser Centre, which comprises a network of national laser-research institutions, was launched at the conference. NEPAD is also backing the upgrading of existing centres, including the International Livestock Research Institute in Nairobi, to which the Canadian government last month committed US\$21 million. ■

Magnetic effect sends physicists into a spin

Geoff Brumfiel, Washington

It was an effect so huge that it promised to increase the storage of computer hard-drives a thousand times over, but physicists are now clashing over whether it was just a mistake.

The effect, called ballistic magnetoresistance (BMR), would theoretically allow manufacturers to boost the data capacity of hard-drives — packing trillions of bits into a single square inch of disk surface.

The promise of BMR edged closer to reality last year, when Harsh Chopra and Susan Hua at the State University of New York at Buffalo revealed that they had observed a massive increase in resistance at the junction between two nickel wires (H. D. Chopra and S. Z. Hua *Phys. Rev. B* **66**, 020403; 2002). Hard-drives depend on similar, less efficient

effects for reading and storing data.

Now, a team led by William Egelhoff of the National Institute of Standards and Technology (NIST) in Gaithersburg, Maryland, says it has results that cast doubt on Chopra and Hua's findings.

BMR occurs when electrons aligned in one direction by an applied magnetic field meet resistance as they try to pass into a wire containing electrons aligned in the opposite direction. To demonstrate the effect, Chopra and Hua used two nickel wires around 100 micrometres in diameter, joined in the shape of a 'T'. But the junction between them was less than a nanometre in diameter — which restricts the electrons' movement and provides the conditions for BMR to take place. With electrons in both wires aligned in the

same direction, current flowed across the junction normally, Chopra says. But when an applied magnetic field caused one of the wire's electrons to change alignment, resistance rose by up to 100-fold.

On 3 November, at the AVS Science and Technology Society's annual meeting in Baltimore, Maryland, Egelhoff suggested that this result may not be BMR after all. He believes that the magnetic field is physically tearing apart the small contact connecting the two wires. "Magnetic fields pushing and pulling on a nanocontact can easily change its shape and resistance," Egelhoff says.

Chopra disputes this, saying that under certain circumstances the resistance drops as the magnetic field increases. He also claims that Egelhoff's set-up is less sophisticated than his own because the junction is too thick. "It's very difficult for me to believe this effect doesn't exist," he says.

A third researcher disagrees with both interpretations. Nicolás García, director of the Spanish Research Council's nanotechnology laboratory in Madrid, says that Egelhoff's analysis of magnetic forces on the junction is incomplete — but that he doesn't agree with Chopra's data either. "The results cannot be reproduced," he asserts. García says that in his experiments, he has consistently obtained BMR that produced a threefold increase in resistance.

This dispute will ultimately have to be resolved by research groups from beyond the small number now working on BMR, says Ami Berkowitz, a physicist at the University of California, San Diego. Detailed study of the connections using more sophisticated equipment is the only way to find out whether the effect is real, he says. ■



The data capacity of computer hard-drives could be greatly increased if a physics result proves true.

Geologists call for desalination of Gaza Strip's water

Betsy Mason, Seattle

Residents of both the Palestinian territory and neighbouring Israel could benefit from action to desalinate groundwater crossing their border, according to a geological study of the Gaza Strip.

A team of Israeli, Palestinian and French scientists funded by the European Union and led by geochemist Avner Vengosh of Ben-Gurion University in Beer Sheva, Israel, announced their findings at the annual meeting of the Geological Society of America in Seattle on 3 November.

Some 1.2 million people in the Gaza Strip depend on water from the southern Mediterranean Coastal aquifer that underlies the territory and extends into Israel. In general, the amount of chloride ions present

in a sample is used as a guide to how much sodium chloride, or salt, the water contains. Drinking-water in the Gaza Strip typically contains more than a gram of chloride per litre, far more than Europe's legal limit of 250 mg or Israel's of 600 mg.

Vengosh's team found that most of the salt comes from naturally saline groundwater flowing into the aquifer from Israel. The researchers then modelled the system and suggested what they say would be a relatively simple solution to the problem. Pumping the saline water from Israel out of the aquifer before it reaches the Gaza Strip and reducing the amount of water drawn from the aquifer in Gaza would lower salt levels, the researchers say. Desalination plants could convert

the salty water to fresh water to make up for the reduction.

"Palestinians will benefit from improved water and Israel will earn goodwill, while only losing saline water it wasn't using anyway," says Vengosh, who estimates that adding just ten wells near the border between Israel and the Gaza Strip and two small desalination facilities could do the job.

But moving the proposal from the scientific arena to the political one will be tricky, says hydrologist Amer Marei of Al-Quds University in Jerusalem. The project will need a sponsor, such as the World Bank, and the political climate in the region will have to improve. "This will not fit with the current political agenda. But that agenda will change," he says. "It will happen." ■



Nuisance neighbours: a number of Gibraltar's macaques have been culled in response to public pressure.

Primatologist rocks Gibraltar by quitting over macaque cull

Quirin Schiermeier, Munich

A top primatologist has pulled out of field research in Gibraltar, in protest at the culling of a group of marauding Barbary macaques that had been part of a long-term conservation study.

Robert Martin, vice-president of the Chicago Field Museum of Natural History, also boycotted an international macaque conference last week in Gibraltar — even though he was one of its organizers.

"The culling without warning of half of our study group has effectively terminated our research project. There is no point in continuing," says Martin.

The couple of hundred semi-wild Barbary macaques (*Macaca sylvanus*) that populate the rock of Gibraltar, a small British crown colony at the southern tip of Spain, are the only non-human primates living outside captivity in Europe. The well-characterized population provides an ideal model for ecologists studying the genetic effects of habitat fragmentation — a key question in primate ecology and conservation biology.

Eight years ago, Martin, then at the University of Zürich, Switzerland, began monitoring reproduction and genetic variability in a 50-strong macaque colony on Gibraltar's Middle Hill. The study was meant to provide guidelines for the conservation management of threatened wild macaque populations in forest fragments in Algeria and Morocco.

It is not known whether the Barbary macaques are indigenous to Gibraltar, but legend says that when they leave, the peninsula will cease to be British. The myth has been taken so seriously that after a critical

decline during the Second World War, Winston Churchill, then Britain's prime minister, issued a 1942 order to import some of the animals from North Africa.

The macaques are a major tourist attraction, but in recent years they have multiplied to nuisance levels. The Gibraltar government has repeatedly ordered the culling of monkeys found roaming built-up areas in search of food. In July, half of Martin's study group were killed after some animals were found vandalizing property and attacking children.

"There was a public-health risk, so we had no choice," says Mark Pizarro, a surgeon at Gibraltar's veterinary clinic.

But Martin says that the monkeys would not have outgrown their habitat if the government had not ignored timely scientific advice. "I submitted as early as 1997 detailed proposals on effective population control by means of contraception, but nothing much was done," he says.

The Gibraltar Ornithological and Natural History Society, which has recently taken over the management of the macaque colony, tried in vain to prevent the culling.

"There was too much public pressure on the government to take action," says John Cortes, general secretary of the society, whose intervention persuaded the government to agree not to have the entire group killed, as had been initially planned.

Martin's withdrawal from the government-sponsored conference is understandable, says Cortes. "But I still believe it would have been better if he had come and made his point here."

■ www.gib.gi/gonhs/gibraltar/congress.htm

Italy's proposal for technology institute riles researchers

Alison Abbott, Munich

Italy has revealed plans to allocate €1 billion (US\$1.2 billion) to set up what the government hopes will become a world-class institute of technology.

But many Italian researchers are suspicious of the surprise proposal, which bypasses existing research institutes and agencies. They also complain that a detailed concept for the facility, named the Italian Institute of Technology (IIT), has yet to be set out.

Cash for the institute was included last month in Italy's budget for 2004, which is likely to be approved by the parliament in Rome later this month.

But the initial plans were first outlined by Vittorio Grilli, the Italian government's chief accountant, at a seminar at Harvard University on 23 October. Grilli, one of the architects of the scheme, told the seminar that Italy needs to improve its track record in research and technology transfer.

The IIT will be funded by an endowment over ten years, with the first €100 million due in January. Italian scientists, however, are less than thrilled that the government plans to lavish funds on a new project while existing institutions are starved of resources.

Italian research has been in crisis for years, as successive governments have tried to implement reforms and cut red tape. Most research organizations are now in the hands of commissioners appointed by the government to oversee such reforms.

Two of the commissioners — Adriano De Maio from the CNR, the national research council, and Carlo Rubbia from ENEA, the energy and environment agency — have spoken out against the IIT concept, arguing that the sort of money set aside for it could have been used to give the reforms a chance to work.

"It's a bit schizophrenic of the government to install commissioners to make research more efficient, then put significant new funds elsewhere," notes physicist Roberto Battiston of the University of Perugia.

Emilio Bizzi, an Italian neuroscientist at the Massachusetts Institute of Technology who has advised successive Italian governments, says he sympathizes with the scientists' concerns. But he adds that he understands the government's frustration with existing institutions. "I'm not surprised the government wanted to bypass the system," he says. "But whether it was wise to do so, I can't say." Government officials declined to comment. ■

United Nations fails to reach agreement on human cloning

New York The United Nations (UN) will wait two more years before making a decision on the contentious issue of regulating human cloning activities.

The decision sidesteps a battle between two competing UN resolutions. One, backed by more than 60 countries including the United States, proposes a blanket ban on both reproductive cloning and therapeutic cloning of human embryos for medical research. The other, supported by more than 20 countries, wants to ban only reproductive cloning.

The UN legal committee instead narrowly passed a resolution proposed by Iran to defer the decision — 80 countries voted in favour, 79 voted against and 15 abstained. In the absence of UN guidelines, countries can continue to regulate human cloning as they choose. The United Kingdom, for example, already bans reproductive cloning but allows therapeutic cloning. In the United States, there are no laws against either practice.

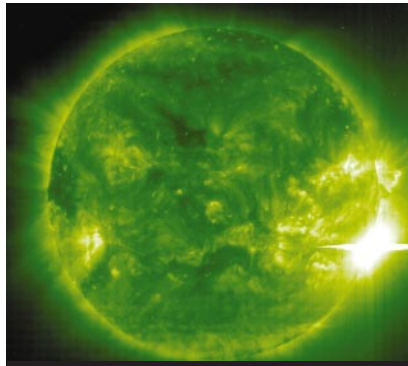
A delay is preferable to a resolution without unanimous support, says Marc Pecsteen, legal adviser at the Belgian mission

to the UN. “We think the outcome is a victory for common sense, but not a satisfactory one,” he says.

The Sun flares up with a record-breaking burst

Washington The Sun belched out a giant solar flare on 4 November — the biggest seen since measurements began in the 1960s.

The solar flare (pictured) hurled a wave of charged particles towards Earth at 8 million kilometres per hour. Although huge, the burst only skirted our planet’s atmosphere and didn’t cause as much of a show as last



Burst of activity: the biggest flare yet recorded follows several active months for the Sun.

month’s barrage of solar activity, which created stunning auroral displays.

The past few months have been some of the most active on record for the Sun, according to Neal Hurlburt of the Lockheed Martin Solar and Astrophysics Laboratory in Palo Alto, California. The activity is confounding solar scientists, because the Sun was thought to be in a quiescent phase of its 11-year activity cycle. “This is supposed to be a boring time,” says Hurlburt.

Emissions row gives EPA head tough start

Washington No sooner was former Utah governor Mike Leavitt sworn in last week as administrator of the US Environmental Protection Agency (EPA) than his agency found itself in hot water.

The *New York Times* reported that EPA lawyers were dropping investigations into some 50 plants that had violated old air-pollution rules but would comply with more lenient rules due to take effect in December. The new rules are part of a move by the Bush administration towards voluntary emission controls.

The agency said a day later that it may still pursue past violators on a case-by-case basis. But Democrats in Congress are still accusing the Bush administration of

pandering to polluters, and are calling for hearings about the case.

Medical tests of cannabis leave doctors in a haze

London A study into the medicinal effects of cannabis looks set to open up a regulatory dilemma. In patients with multiple sclerosis, cannabis has no effect on spasticity — a condition where tight muscles lead to uncontrollable jerking motions — according to objective measures of muscle function. But two-thirds of patients treated with cannabinoids felt as if their spasticity had improved. These patients didn't notice any

effects on their mood or their energy levels.

Cannabis is not licensed for medicinal use in Britain, but many patients smoke it illegally anyway. So the UK Medical Research Council conducted a large, placebo-controlled study of cannabis extracts and synthetic samples of the drug's most active ingredient — tetrahydrocannabinol (THC) — to test anecdotal claims of symptom relief. The trials involved 630 patients at 33 UK centres.

Medics say that the patient perspective is important, but it is now up to the licensing authorities to decide what to do.

View of space expands in both directions

Tokyo Astronomers broke records for both the closest and the most distant galaxies ever spotted last week, as two sets of researchers peered into very different regions of space.

A group of French, British, Australian and Italian researchers announced the discovery of the Canis Major dwarf galaxy, a mere 42,000 light years from the centre of the Milky Way. The group used the Two-Micron All Sky Survey — a pair of infrared telescopes based in Arizona and Chile that can peer through dust in the Milky Way that obscures astronomers' view of nearby objects. Canis Major appears to be vanishing as it is sucked into our Galaxy.

Researchers at the National Astronomical Observatory of Japan then reported that their Subaru telescope in Hawaii had found a 13.7-billion-year-old galaxy 12.8 billion light years away. Subaru has found eight of the nine most distant known galaxies.

Plans to rename Livermore laboratory fall out of favour

Washington Lawmakers have halted a plan to rename the Lawrence Livermore National Laboratory after hydrogen-bomb pioneer Edward Teller, who did much of his work at the California lab.

After Teller's death in September (see *Nature* 425, 226; 2003), Duncan Hunter (Republican, California), who chairs the House Committee on Armed Services, proposed legislation to rename the laboratory. In an unusual alliance, anti-nuclear activists sided with the Republican, saying that the name was perfectly appropriate for a weapons lab.

But lab officials and Congresswoman Ellen Tauscher (Democrat, California) were less keen to name the lab after such a controversial figure. The proposal has now been scrapped from this year's defence authorization bill. Tauscher says that the lab is considering giving Teller's name to a new US\$92-million supercomputing facility.



Joint decision: many multiple-sclerosis patients believe that smoking cannabis brings relief.



Their lives in his hands: athletes' reputations can be shattered if Don Catlin finds a banned drug in a sample.

No dope

Don Catlin's lab has struck a major blow against drug abuse in athletics, by developing a test for a shadowy 'designer steroid'. Jonathan Knight visits the scientists who are striving to keep sport clean.

When you arrive at the Olympic Analytical Laboratory in Los Angeles, it's clear that this isn't an operation that goes out of its way to make visitors feel welcome. Nestled among car-body shops, its narrow façade shoulders up against chain-link fences topped with razor wire. The mirrored-glass front door of this drug-testing lab is locked, and will only be opened if you are meant to be there. Numerous television crews have been turned away in recent weeks, and a sign inside the door reminds staff not to admit an athlete who filed a complaint against the lab after it found steroids in her urine.

Although the fences actually belong to the neighbouring car-repair yards, this defensive appearance is fitting for a lab that is at war with the suppliers of performance-enhancing drugs. After a summer of intense work on the residue from an anonymously supplied syringe, the lab rocked the world of sport last month by revealing that an undisclosed number of athletes had tested positive for a designer steroid, known as tetrahydrogestrinone or THG, which had evaded standard tests (see *Nature* 425, 752; 2003).

Keeping things under wraps is an essential part of the game, says lab director Don Catlin. Indeed, most of the lab's employees were in

the dark about THG until after the news became public. "If what we were working on had gotten out, people would have stopped using it before we had a test," Catlin says.

Drug testers and drug-using athletes have been locked in battle for more than three decades. When the International Olympic Committee (IOC) ran its first drug tests during the 1968 games in Mexico City, no one was found using anything other than alcohol. But heavier testing in Munich in 1972 caught seven athletes using banned drugs. Since then, testers and dopers have been engaged in an arms race — with the users of illicit drugs generally being perceived to have the upper hand.

An unlikely hero

Catlin did not expect to become involved in this struggle. In 1982, he was an established endocrinologist at the University of California, Los Angeles (UCLA), when the IOC came to ask if he would run the testing for the 1984 Los Angeles games. "They showed me a list of banned substances," Catlin recalls, "and I said, 'I don't think we can do this.'" But when the IOC made it clear that it was offering to greatly expand his lab and hire the staff he would need, he rose to the challenge.

Today Catlin's lab, which is still part of

UCLA, has three dozen employees. Its biggest customers, apart from the US Anti-Doping Agency (USADA), are the National Collegiate Athletic Association and the National Football League. Most of the work involves the routine processing of number-coded urine samples that arrive by courier almost daily. These can be analysed for several dozen hormones, stimulants, diuretics and other banned substances, depending on what the customer asks for. Most of the tests involve a combination of gas chromatography and mass spectrometry. The desktop machines that run these procedures occupy almost every available surface in the lab.

Besides this routine, there is a continuous effort to improve the tests. In the EPO lab, where technicians run antibody tests to detect the hormone erythropoietin, which boosts red-blood-cell counts so that more oxygen gets to the muscles, group leader Andreas Breidbach is looking for ways to see further back in time. Right now, evidence of EPO use is hard to find after about a week. But the benefits last longer. Knowing this, many athletes stop doping ahead of an event. Breidbach hopes to extend the window of detection by tweaking each step of the test to improve its sensitivity. "We are always improving because the athletes are always improving," he says.

One of the lab's research achievements is a method for spotting testosterone abuse. Unlike synthetic steroids, testosterone and its breakdown products are naturally present in the urine of dopers and non-dopers, whether they are male or female. The standard test for testosterone doping looks for an elevated ratio of the hormone relative to its cousin epitestosterone. These are normally present at equal levels in the urine, and the IOC has set a ratio of 6:1 as the cut-off for declaring a positive test. This is bad news for the rare athlete with a naturally elevated ratio. And the test is not hard to beat: athletes can, for instance, take epitestosterone as well as testosterone.

Catlin's group hit on the idea of using carbon isotope ratios to distinguish between natural testosterone and its lab-made counterpart. Synthetic testosterone contains less of the heavy isotope carbon-13 compared with the natural compound. This is because the hormone's manufacturers begin their synthesis with an extract from yams, which happens to contain a compound with the same four-ring structure as steroids. Plants incorporate carbon-13—present at low levels in the atmosphere—much less efficiently than do animals. Although the isotope test is expensive and not always used, it has made a difference. Once athletes became aware of the test in the late 1990s, many stopped using testosterone, Catlin says.

Peak performance

But the battle continues. Just how far ahead are drug-using athletes and their suppliers? No one knows for sure, although part of the answer may lie in a slim binder tucked away on a shelf in the spectral analysis room at Catlin's lab.

In this room, Yuliya Kucheroва pores over gas-chromatography traces and mass spectrometers, looking for the signatures of steroids. In some cases, the steroid shows up as a distinct peak by gas chromatography, which separates compounds based on their volatility. In others, confirmation by mass spectrometry is essential. This involves breaking molecules into fragments and determining their molecular weights.

Kucheroва, after reading dozens of spectra a day for years, has no trouble spotting banned compounds. But occasionally a new peak shows up that doesn't correspond to any known drug, prescription or otherwise. Is it a new synthetic steroid? A Chinese herb? A rare prescription medication? These mystery peaks wind up in the special binder, and several dozen are awaiting further investigation.

Once in a while, Catlin and his colleagues get lucky. Kucheroва last year spotted a peak she had actually seen previously. It was the synthetic steroid norbolethone, first made in 1966 as a possible treatment for short stature. But it proved too toxic and was scrapped by its maker, the pharmaceutical

company Wyeth. It has not been made commercially since.

Although only one athlete tested positive for norbolethone, that result was a turning point for the lab. It helped Catlin to convince the sporting authorities that the underground market for steroids was being served by clandestine chemists and manufacturers. He made the case that to keep ahead of the game, the testers needed to work in secret on what the next underground drug might be. Earlier this year, the USADA awarded Catlin's lab a grant to probe such questions.

The group was preparing to launch into this work when the anonymously posted syringe came through the door. A team of three, which rapidly rose in number to eight, was immediately assigned to the task of identifying the residue inside. People worked late, sometimes all night. Details of the



By the numbers: identified only by a code, rows of athletes' urine samples are screened for a wide range of banned substances.



investigation were shared internally on a 'need-to-know' basis, but the sense of urgency and excitement was pervasive. "We've never had a project of this magnitude," enthuses Michael Sekera, the lab's scientific director. "It was an amazing experience."

It was clear right away that normal tests would never have spotted THG. Before analysis, samples are chemically treated to make steroids show up as single sharp peaks in the gas chromatograph. But this process sees THG produce 25 little peaks that don't look like a steroid at all.

Urine trouble

The researchers turned to mass spectrometry to get a clean signature. Then they worked backwards to guess at the compound's structure. They confirmed their hunch by making the compound from scratch and showing that it produced the same signature pattern. Then they found a primate laboratory that was willing to give THG to a baboon, so that they could have some idea how it would come out in the urine. Catlin and his colleagues devised a new test, and began running it on athletes' urine samples. The whole process took two months.

The story isn't over, Catlin says, as he anticipates that there may be legal challenges. These have become more frequent as the stakes in sports have risen. When Catlin first started in the drug-testing business, athletes often reacted to being caught with resignation, even contrition. "You caught me doc, I messed up," Catlin recalls one athlete saying. Now, athletes' lawyers send chemists into the lab in the hope of finding flaws in its testing procedures.

So Catlin has to be very careful. He won't discuss the results of the THG tests as long as the federal investigation into the scandal continues. Not that he knows the names of the accused athletes. All samples are coded before they arrive. Even an athlete's attorneys, who are entitled to observe the testing, say only which number they represent, and not which athlete.

In any case, Catlin and Sekera say that they are more interested in where to go from here. To keep ahead of the people making drugs you have to start making them yourself, they argue. You need to design them, find out how they break down in the body, and develop a test for them.

Whether the lab will get the funds to continue with this approach is a matter for the sports authorities to decide. The bulk of the USADA grant was consumed in figuring out how to test for THG. But if Catlin and his colleagues do resume this work, don't expect them to breathe a word of what they're up to. "We are very good at keeping things to ourselves," Catlin says. ■

Jonathan Knight is a contributing correspondent for *Nature*.

Night club

Growing numbers of amateurs are getting serious about astronomy. The professionals applaud their enthusiasm and success in collecting data and building telescopes — as long as they don't start competing with them for funding. Geoff Brumfiel joins the graveyard shift.

“Growing up, I was one of those guys who wanted to be a rock star,” says Michael Koppelman. Shortly after he graduated from Boston's Berklee College of Music in 1988, the soft-spoken, hard-rocking Koppelman began a ten-year pursuit of his dream — working in recording studios with artists such as Prince and Booker T. and the MGs.

But his life took an unexpected turn when his girlfriend gave him a telescope for his birthday a few years ago. “I got the bug real bad,” he says. He began by imaging stars and galaxies, but gradually became interested in stars whose brightness changes over time. To monitor such ‘variable stars’ precisely required advanced equipment and expertise that Koppelman didn't have, so he began cruising Internet chat rooms, talking to others who were interested in doing more than just taking pictures of galaxies and planets. They were a motley bunch — from dotcom entrepreneurs to construction-equipment salesmen — whose passion for astronomy drove them to spend thousands of dollars on telescopes and cameras.

Extreme hobbyists can be found in any field, but with the help of cheaper technology, such as digital cameras and powerful personal computers, these amateur astronomers are beginning to take measurements that can rival those of the professionals. And the experts are taking notice: astronomers have begun tapping amateurs for their talent and telescopes, using them to detect asteroids, colliding stars and bursts of high-energy γ -rays from distant galaxies. Some amateurs are even turning semi-professional, offering their equipment



Garden variety: the amateur-run Tenagra Observatory has provided data for professional stargazers.

and services to full-time researchers for a price (see ‘Home help’, opposite). “The distinction between amateur and professional astronomers has become very fuzzy,” says Arne Henden, a professional astronomer at the US Naval Observatory in Flagstaff, Arizona.

Amateur army

Across the world, thousands of amateur astronomers point their telescopes skywards each night. Most are sightseers, scanning for colourful clouds of gas or spectacular spiral galaxies, but a few are interested in the more technically challenging aspects of astronomy. They join groups such as the American Association of Variable Star Observers, which since 1911 has enlisted the help of volunteers to monitor variable stars. There are tens of thousands of these stars — too many for professional astronomers to observe regularly — and logging their fluctuations is a full-time operation. Variable stars change brightness for many reasons:

they may pulsate or be eclipsed by a companion star, for example. For the brighter stars, all that an association member needs is a telescope and a reference star with which they can visually compare changes in brightness over several nights.

A small fraction of these amateurs take their measurements a step further. They buy advanced digital cameras and computer software that lets them count the light particles coming from the star at any given time. These more detailed measurements are exactly the type of data that Joe Patterson, an astronomer at Columbia University in New York, is looking for. Since the 1980s, Patterson has studied star systems in which a normal star circles around, and is sucked into, a much heavier object such as a black hole. Like variable stars, these star systems brighten and dim over the course of each orbit, but unlike most single stars, the high speed of the orbit means that the changes can take place in a matter of hours. This means that round-the-clock observations are needed to watch the star as it spins ever closer to its demise.

That's where the amateurs come in, says Patterson. Since the early 1990s he has enlisted the help of enthusiasts around the world to monitor these star systems 24 hours a day. It started with a few well-placed observers on the opposite side of the world from him, but gradually more have signed on, and Patterson's group — the Center for Backyard Astrophysics — now boasts some 40 amateurs from Utah to the Ukraine.

Among the centre's most prolific members is Berto Monard, based in Pretoria,



Key change: musician Michael Koppelman wants to swap chords for constellations.



South Africa. A metrologist at South Africa's Council for Scientific and Industrial Research, Monard became interested in stargazing when a friend at work introduced him to the hobby of spotting satellites as they sail by. "After a few evenings I got a bit bored with the satellites and started watching variable stars," Monard says. His obsession with precision measurements led him to take on even greater challenges, such as monitoring γ -ray bursts — short-lived emissions of high-energy radiation that are thought to be the violent explosions of dying stars. Such bursts of invisible γ -rays last for only seconds or minutes, although their visible 'afterglow' can last for hours, so speed and precision are of the essence, and a special challenge for the amateur connoisseur of cosmic events.

Home help

Some amateur stargazers are helping the professionals to cut the cost of building and running telescopes by offering them the use of home-made alternatives.

Tim Puckett, a construction entrepreneur turned amateur astronomer, has set up a company that builds 0.5-metre telescopes for a fraction of the retail cost. The design is based on Puckett's first telescope, which he built from scratch, and he has a patent pending on part of the design. Puckett, who already has clients in Canada, Italy and Ireland, says that this type of telescope retails for around US\$400,000, but boasts that he is "making the same 'scope for around US\$80,000".

Michael Schwartz, a retired software

Monard's observations of γ -ray bursts have won him international recognition. In July, while professional astronomers mingled at a conference in Australia, Monard was the first amateur to locate the afterglow of a γ -ray burst. Only a month earlier, he appeared as an author on a *Nature* paper detailing the complex structure of such afterglows (M. Uemura *et al.* *Nature* **423**, 843–844; 2003).

Making the grade

Monard's success shows that amateurs can gather top-grade data, but they are still outside the professional community — and that can cause problems. "I've been turned away by some professionals," says Tim Puckett, a retired construction-equipment salesman and amateur astronomer in Atlanta, Georgia. Puckett and his team search for supernovae, the explosions of dying stars. "They say that every supernova we find eats into their grant money," he quips. But in some cases, amateurs and professionals are now banding together to apply for funding. This sort of direct competition can cause friction, says Patterson, whose Center for Backyard Astrophysics has occasionally sought support for group members in developing nations to buy equipment.

For some professional astronomers in poorer countries, buying amateur-standard equipment is the only way that they can do any research. At the Torun Centre for Astronomy in Poland, astronomer Andrzej Niedzielski and his graduate student Gracjan Maciejewski are using a small telescope and a wide-angle camera to conduct a sky-wide search for variable stars. "To analyse the data we use professional techniques, but the hardware is strictly amateur," Niedzielski says. He also enlists a combination of students and amateurs (usually friends brought along for the evening) to operate his telescope. "You might call it a cheap source of labour, but these people like

observing," he says. Niedzielski hopes that eventually 10 or 20 stations around the globe can mount a massive all-sky search for variable stars.

Looking to the future, professional astronomers are planning ever more ambitious projects for their amateur cohorts. Tim Castellano, a planetary scientist at NASA's Ames Research Center in Moffett Field, California, hopes to use amateurs to spot Jupiter-sized planets crossing in front of distant stars. Observing these far-off eclipses requires exceptionally stable detectors and a great deal of skill, and professional astronomers have themselves only spotted one or two such events so far. Castellano believes there are only about 10 or 20 amateurs in the United States who have the right equipment and are up to the task. Training them will require lengthy one-on-one tuition, but Castellano believes the possibility of spotting many new planets makes it worth the effort. "The prize is large," he says.

Hunting planets and publishing papers sounds appealing, but there are limits to what amateurs can do and how fast they can do it. The Amateur Sky Survey, led by Tom Droege, a retired particle-accelerator engineer based in Batavia, Illinois, is a case in point. The project's ambitious goal is to use three digital cameras that are four-megapixels each in resolution to photograph the entire sky each night. But the group has struggled to find amateurs willing to write the software needed to turn images into hard data. As a result, the project has advanced more slowly than its collaborators would have liked, says Arne Henden, who works closely with the group. But the team has nevertheless managed to get almost a terabyte of data, Henden says — enough to fill 1,300 compact discs.

A more fundamental problem is that many amateurs are motivated by things other than science. The technical challenge of taking the data can be more exciting than the statistical analysis and theoretical interpretation needed to turn measurements into scientific papers. "I am not personally concerned with the academics," admits amateur Michael Schwartz of Patagonia, Arizona, Patagonia, Arizona, who runs the Tenagra observatory (see box).

But Koppelman, the reformed rock star, is interested in science. He has enrolled at the University of Minnesota in Minneapolis in the hope of gaining a PhD in astronomy. He knows he has a lot to learn, admitting that "I didn't even know what calculus was", but he is sure that he has what it takes to become a professional astronomer.

Koppelman isn't letting his ambitions get the better of him: while he pursues heady dreams of fame and fortune in professional astronomy, he hasn't given up the day job. "I still play guitar in a band or two," he says. ■

Geoff Brumfiel is *Nature's* Washington physical sciences correspondent.

developer, who made his fortune creating data-encryption programs for automated banking machines, has spent hundreds of thousands of dollars and countless hours putting together an 0.8-metre telescope complete with automated tracking and data-management systems.

Schwartz has begun selling telescope time to the professional community "just to see what happens". So far his customers are mainly based in the United States, but Schwartz believes that interest in his service could grow, especially in countries where researchers lack access to reliable medium-sized telescopes.

► www.observatoryscope.com

► www.tenagraobservatories.com

How genius can smooth the road to publication

If at first your paper doesn't succeed, try, try — and try to find a brilliant supporter.

Sir — In the entertaining Editorial on journals' *faux pas* ("Coping with peer rejection", *Nature* **425**, 645; 2003) the reader is referred to an undoubted success, "perhaps the most celebrated editorial judgements of all" — the publication in *Annalen der Physik* of five extraordinary papers, all written by Einstein in 1905.

However, when reconciling these events with the current peer-review system, it is worth noting that none of Einstein's papers were sent to reviewers. The decision to publish was made exclusively by either the editor in chief, Max Planck, or the co-editor, Wilhelm Wien — both 'peers' beyond doubt who were later to win the Nobel prize in physics. The importance of these editorial judgements is underlined by the decision of UNESCO to declare 2005 the World Year of Physics to celebrate the centenary of Einstein's 'miraculous' year.

This year we celebrated the 50th anniversary of the publication of the paper by Watson and Crick describing the structure of DNA (*Nature* **171**, 737–738; 1953). Likewise — according to John

Maddox, a former editor of *Nature*, quoted in the *New York Times* (25 February 2003) — this manuscript was never sent out to reviewers. The editors accepted the paper upon receipt of a "Publish" covering letter from Nobel laureate Sir Lawrence Bragg.

Thus, to complete the final moral for rejected authors of presumed Nobel-winning work — persist, and get in contact with a noble genius.

Jens Brümmer

*Institut für Klinische Chemie,
Universitätsklinikum Hamburg-Eppendorf,
20246 Hamburg, Germany*

John Maddox replies — As Jens Brümmer states, the Watson and Crick paper was not peer-reviewed by *Nature*. I have two comments on this. First, the Crick and Watson paper could not have been refereed: its correctness is self-evident. No referee working in the field (Linus Pauling?) could have kept his mouth shut once he saw the structure. Second, it would have been entirely consistent with my predecessor L. J. F. Brimble's way of working that Bragg's commendation

should have counted as a referee's approval. Brimble, who used to "take luncheon" at the Athenaeum in London most days, preferred to carry a bundle of manuscripts with him in the pocket of his greatcoat and pass them round among his chums "taking coffee" in the drawing-room after lunch. I set up a more systematic way of doing the job when I became editor in April 1966.

An interesting question is how Wilkins and Franklin came to have papers in the same issue. J. T. Randall, the head of physics at King's College, was a pal of Brimble's co-editor at the time, A. J. V. Gale, and it is known that Crick sent a copy of his paper to Maurice Wilkins. (He probably sent one to Rosalind Franklin as well.) My guess is that Randall would have called up Gale the minute he heard about Crick's paper, pleading for equal treatment for King's. There's a letter from Wilkins to Crick saying: "Franklin has jumped on the bandwagon — Christ!"

*John Maddox, now Emeritus Editor, was
Editor of Nature during the periods
1966–1973 and 1980–1995.*

Phosphorus: time for us to oust bad spelling

Sir — I have long suspected that "phosphorus" is the most frequently misspelt word in the environmental sciences, with "phosphorous" — the adjectival spelling — being the primary offender. Although phosphorous is a legitimate adjective meaning phosphorus-rich, this is rarely the meaning intended when this spelling is used in the literature. Your News story on the politics of phosphorus pollution in the Everglades, "Judge's sacking rocks Everglades clean-up" (*Nature* **425**, 551; 2003), provides yet another example of this confusion. The word appears eight times in the News story and photo caption, three times ending in "-ous" and five times "-us" — the ratio being in favour of propriety, but only marginally.

Many students and colleagues who misspell the word assert that phosphorous is the British spelling. This is clearly not the case. But I wonder if *Nature*, as a journal with British spellings but a wide audience in the United States, intended a compromise solution to this spelling problem in its News story.

Perhaps the idea is that "yous guys" (an Americanism) and "us guys" can compromise and agree that either

spelling is acceptable. But I rather think it is just "us".

Nelson Hairston Jr

*Department of Ecology and Evolutionary Biology,
Corson Hall, Cornell University, Ithaca, New York
14853-2701, USA*

Drinking your health? It's too early to say

Sir — There has been much debate in the media about the potential health benefits of moderate wine consumption, and the components of wine that may be responsible. Recently, Konrad Howitz and colleagues reported an increased lifespan for yeast treated with resveratrol through the activation of sirtuins (*Nature* **425**, 191–196; 2003). Because resveratrol is a polyphenol found in wine, these findings have been linked in media reports to studies suggesting that regular wine drinking can increase longevity (for example, see www.nature.com/nsu/030825/030825-1.html; and S. S. Hall, *Science* **301**, 1165; 2003).

However, the claims being made for resveratrol are unjustified. There is no evidence that it would provide these benefits as part of a normal diet — even for wine drinkers.

It is not a simple matter to extrapolate

the results of yeast studies to human health, especially when the studies on yeast used much higher concentrations of resveratrol than are available from wine drinking. Levels of resveratrol in wine are generally less than 5 mg per l, and it is heavily metabolized during the absorption process, resulting in extremely low plasma concentrations (D. M. Goldberg *et al.* *Clin. Biochem.* **36**, 79–87; 2003).

In addition, the link between regular wine consumption and longevity is still not proven. Deaths from coronary heart disease and from all other causes are lower in people drinking two to three glasses of wine a day. This has been interpreted as evidence that regular wine drinking could increase longevity, but it has not yet been confirmed by population studies showing, for example, a large increase in the number of people who live beyond 80 years.

Wine drinking in the pursuit of a longer life should not be encouraged until we have a better understanding of the link between diet and health.

Roger Corder*, Alan Crozier†, Paul A. Kroon‡

**William Harvey Research Institute, Charterhouse Square, London EC1M 6BQ, UK*

†Plant Products and Human Nutrition Group, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

‡Institute of Food Research, Colney Lane, Norwich NR4 7UA, UK

Web masters

Spiders' silky skills hold the key to their evolutionary success.

Spider Webs and Silks: Tracing Evolution from Molecules to Genes to Phenotypes

by Catherine L. Craig

Oxford University Press: 2003. 256 pp. \$59.95

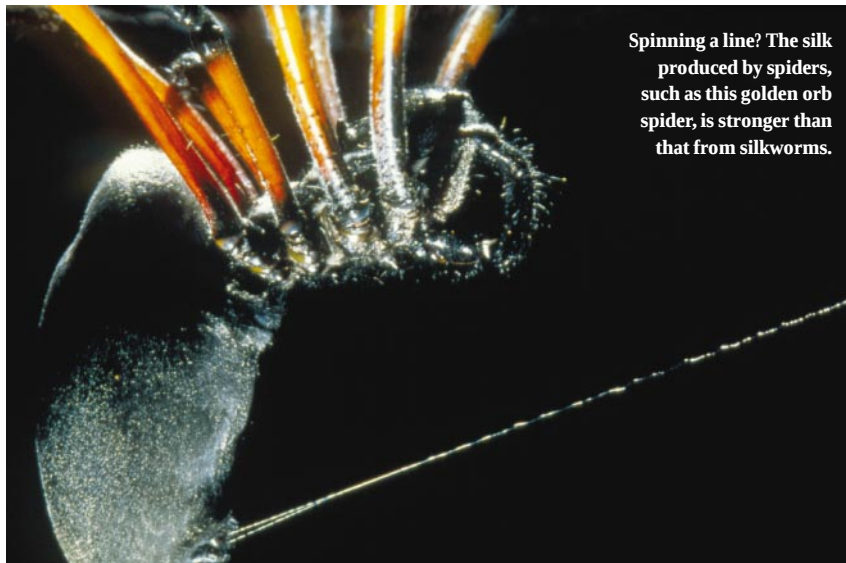
Fritz Vollrath

Spiders are masters of the extended phenotype. The archetypal orb web spider, for example, manages to expand its effective body size by at least an order of magnitude. This remarkable extension is superbly matched to the local conditions, and total daily restoration is often thrown in for free. How does the spider perform this clever trick of morphological inflation? By building its trademark web; by matching and redesigning the structure to suit the changing circumstances of location, weather and prey; by rebuilding it daily; by recycling the silky hardware, so it can devote most of its resources to the software and the 'running costs' of building. Running, or rather walking, is the operative word here, as the web is a frozen record of the spider's path and its manual labour, written in silk.

Silk is the key to the spider's success, with behaviour playing merely a supporting role. Without their silk, spiders would be weaklings among their arthropod peers — if they had survived at all. They are soft-bodied and so are prone to physical damage; they breathe through lungs, and are in constant need of high humidity; and they are wingless, and their hydraulically driven legs buckle when overused. As a result, spiders would have been no match for the tough and virtually indestructible insects. But silk gives spiders a distinct edge, and has become the main weapon in an arms race with their insect relatives, which are both their main prey and major predators.

Mention silk to a polymer chemist and they will get dreamy eyes. It is a natural fibre to match the best man-made ones, but its production is as eco-friendly as it gets, occurring at ambient temperature and near-ambient pressure, with water as the solvent. How is it done? We don't know. Why is it so tough? We don't know. Actually, what is silk, exactly? We don't know. So there's lots of scope for research, then.

This is especially gratifying because state-of-the-art analytical techniques such as X-ray diffraction, Raman spectroscopy and mechanical testing have recently been refined to allow measurements on single silk filaments, which are often less than 2 μm in diameter. Today, single-filament analysis can be done online as the fibres are being spun by the animal; the speed at which the spider



Spinning a line? The silk produced by spiders, such as this golden orb spider, is stronger than that from silkworms.

P. GOETGHELUCK/SPL

reels the silk can be altered over five orders of magnitude by heating and cooling the spider.

Although we have no real answers to even general questions, we have come a long way in our understanding of silks, and Catherine L. Craig's *Spider Webs and Silks* is an extremely useful staging post. I don't agree with everything that Craig has written, but even so I can highly recommend the book as a comprehensive and up-to-date account of silk.

The four chapters on silk form the core of the book. The first outlines the history of silk evolution, and the second explores the genetic code behind the proteins that make up some silks. Next comes an investigation of how the mechanical properties of bulk silk depend on the protein skeleton of the filament. And it is all wrapped up by an explanation of the economics of silk synthesis and its effect on the evolution of the wide range of silk types that some spiders can produce.

There are also three intriguing chapters on spider webs and their role in feeding the beast, as well as in constraining the more social aspects of its evolution. These chapters are not so much a review as a personal voyage, relying heavily on the author's own work, which focuses on the interaction of insect vision and a web's architecture as well as the silk's structural properties. There is also a chapter on the absence of higher social development in spiders, which, in a twist, is linked to development and related silk-production costs. The brief final chapter summarizes the author's view on the forces that drive silk evolution.

The writing throughout is clear and well presented, and even though there is no glossary, an effort has been made to avoid jargon.

There is a good index and the references are, on the whole, comprehensive, although with some curious gaps and biases. Overall, the book provides excellent value for money on a number of levels.

Silk research has undergone something of a renaissance, with the focus of study shifting from the classic (and highly commercial) silkworm silk to the more esoteric spider silk. Silkworms spin through their mouth between their nippers, which means that they can cut the fibre at will. Most silkworm silk is collected from preformed cocoons that the worm has spun at its leisure, and it invariably contains numerous weak points resulting from the worm's typical spinning action. By contrast, spiders spin from their bottoms and cannot interfere with forced silking as long as their legs are kept clear, so spider silk is typically collected as it is spun under highly controlled conditions.

It may be that spinning conditions, rather than silk genes, are a major factor for fibre quality, as silkworm silk that is spun spider-fashion can be almost as tough as spider silk (Z. Shao & F. Vollrath *Nature* **418**, 741; 2002). Using spider silks produced under controlled conditions allows the construction and testing of hypothetical links between protein form and folding. In this way, spider silks provide a new way of studying fibrous proteins. For example, recent studies have shown that spider silk shares important protein-folding configurations with amyloids and prions (J. M. Kenney *et al. Eur. J. Biochem.* **269**, 4159–4163; 2002), which suggests that unravelling the spider's way of making a tough fibre might have surprising results.

Craig's book has the somewhat puzzling

subtitle "Tracing evolution from molecules to genes to phenotypes". Most authors would have gone from genes to phenotypes without bothering with the molecular level. But not Craig, who states that one of her goals is "to illustrate the ease with which evolutionary studies of spiders and silk proteins can cross traditional molecular and organismal borders". Thus silk proteins are "the perfect system through which to study evolutionary conflicts among molecular genetic constraint, protein architectural constraint, protein diversity and selection". Indeed, silk has evolved to function as a dead material outside the animal's body. And silk is only one step (spinning) removed from the protein in its native form; the chain from gene to raw protein to dehydrated protein fibre is brief.

But more importantly, the fibre, once drawn — whether as a simple drag safety line or as a structural member in a web — is the phenotype that is selected more or less on its own, distinct from its creator. This is unusual for a protein and suggests that silk might allow for shortcuts in the study of protein evolution — yet another reason why studying silk might provide interesting insights into the more general aspects of protein folding.

Be that as it may, Craig's *Spider Webs and Silks* brings a fascinating and important subject to a potentially broad audience. And it might even turn some arachnophobes into arachnophiles. ■

Fritz Vollrath is in the Department of Zoology, South Parks Road, University of Oxford, Oxford OX1 3PS, UK.

From genes to biochemistry

George Beadle, an Uncommon Farmer: The Emergence of Genetics in the 20th Century by Paul Berg & Maxine Singer
Cold Spring Harbor Laboratory Press: 2003.
383pp. \$35, £25

Benno Müller-Hill

Genetics is a young science, having its roots just 100 or so years ago in the rediscovery of Gregor Mendel's work on inheritance in peas. These days, this history is left out of most textbooks and is rarely taught at university. Today's students learn only about what has happened in the past five years, not about the pioneering work that opened up their field — they know nothing of the giants on whose shoulders they stand. If you ask today's advanced genetics students about George Beadle (1903–1989), you would be met with a deafening silence. So it is fortunate that Paul Berg and Maxine Singer took the time to write a book about this great pioneer.



Field work: at Cornell, George Beadle (kneeling) studied maize genetics with Barbara McClintock (right).

They describe Beadle's family and his childhood and youth on a farm close to Wahoo, Nebraska. His father prevented his elder brother from going to agricultural college, but when this brother died after being kicked by a horse on the farm, George was allowed to go the agricultural college in Lincoln, where he developed an interest in maize genetics. His intelligence and hard work impressed his teachers, who sent him to Cornell University in Ithaca, New York, where he continued to work on maize genetics. There he met Barbara McClintock, who demonstrated that maize has ten chromosomes. But her supervisor was not amused when she published this under her own name without including him as a co-author.

Beadle's work on maize genetics earned him a PhD in 1929, after which he moved to the California Institute of Technology, which was an exciting place to be. Thomas Hunt Morgan had just moved there, and the authors describe his fly group in a separate chapter that is a must for students. I would have liked to hear more. For example, the authors mention that the undergraduate Alfred Sturtevant produced the first linear map of five genes of the fruitfly's X chromosome, and that Morgan allowed him to publish his results without demanding his own name on the paper. Berg and Singer don't comment on this, but how many professors would allow this today?

Beadle started collaborating with Boris Ephrussi, who was on sabbatical there, to try and understand what a gene produces. For Morgan and his group, the gene was abstract, but Beadle and Ephrussi wanted to come closer to biochemistry. They concentrated on *Drosophila* mutants with different eye colours, *vermillion* and *cinnabar*. After five years they concluded that tryptophan is converted by the *vermillion* gene into a substance of unknown structure. So it was a blow in 1940 when they read in the journal *Naturwissenschaften* that Adolf Butenandt had identified the compound as 3-hydroxykynurenine. Five years' work and then scooped!

Ephrussi then moved into a different field but Beadle stuck to the problem. During a lecture by his postdoc Edward Tatum, Beadle had the idea of treating micro-organisms with mutagens and isolating mutants that are unable to produce amino acids, hormones or other substances. These mutants will then show that one gene produces one enzyme. Tatum did the experiment using the fungus *Neurospora*. It worked like a charm, and hundreds of such mutants were isolated and analysed. What worked well with *Neurospora* worked even better with the bacterium *Escherichia coli*. One of Tatum's students, Joshua Leder-

berg, showed that such mutants could be isolated, and demonstrated genetic exchange between two mutants of *E. coli*. In 1958, Beadle, Lederberg and Tatum shared a Nobel prize for their work. As Ephrussi wrote to a friend: "I suddenly felt my life wasted."

But the story does not end here. When the experimental problem was solved and Beadle had convinced the sceptics, he moved into the field of administration at Caltech, first as chairman of the biological faculty and then as president. He excelled at hiring excellent faculty and defended Linus Pauling when he was attacked as a communist. Even after his retirement, Beadle returned to a familiar problem: was teosinte, a wild grass from Mexico, really the predecessor of maize?

The book tells us in detail about Beadle's two marriages, the salaries he earned (but not their equivalent values today), his journeys by ship and by train, and the fact that he succumbed to Alzheimer's disease. There is plenty here for everyone. Those interested in the history of genetics will want to read the whole book, but today's students would benefit from just a few chapters. ■

Benno Müller-Hill is at the Institute of Genetics, Cologne University, Weyertal 121, 50931 Cologne, Germany.

A recipe for success?

Organic Syntheses Database
Wiley/Organic Syntheses, Inc: 2003.
<http://www.interscience.wiley.com/db/oss>
Available through a site licence.

John Mann

The total synthesis of complex natural products remains one of the most tangible and often elegant demonstrations of the chemist's craft. Some practitioners have turned these endeavours into something approaching an art form. It is also often

Installation

Making haze

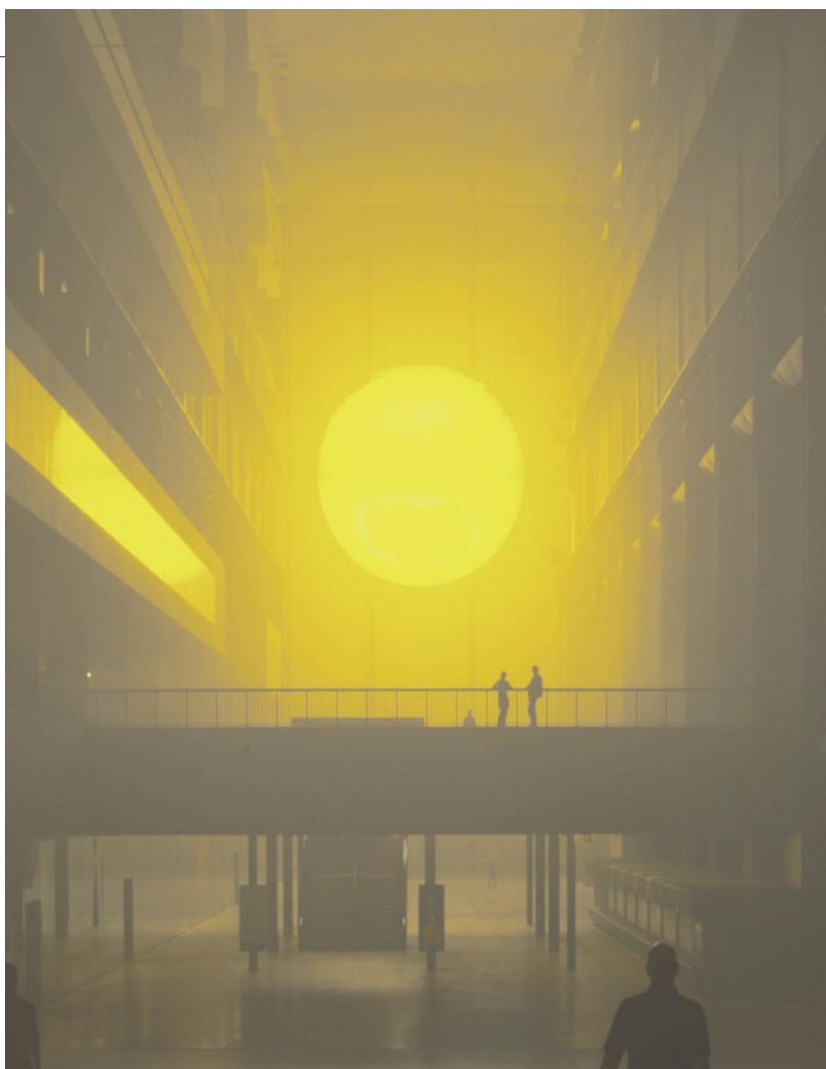
Built specifically for the vast, multistorey space of the Turbine Hall at Tate Modern in London, Olafur Eliasson's installation *The Weather Project* is mesmerizing. A huge sun blazes at the far end of the hall, its light diffused by artificial fog, creating a dynamic microclimate.

The Danish-Icelandic artist's work investigates the dichotomy between nature and culture. Here he tackles the weather, a ubiquitous topic of conversation, which he says "somehow defines a sense of community, within which conflicting views can be accommodated, such as liking or not liking rain". By introducing meteorological variables — water, light, temperature and pressure — into artificial, built environments, he encourages viewers to reflect on their understanding of the natural world and their relationship to it. Eliasson is intrigued by perception and describes the moment when people pause to consider what they are experiencing as "seeing yourself sensing". But even though all is not as not seems, his aim is to heighten people's awareness of their perception, not to deceive them.

At ground level, visitors can clearly see machines pumping artificial haze into the hall and the bank of mono-frequency lamps that powers the semicircular 'sun'. From above, they can observe that the reflective 'sky' is suspended from the ceiling, both doubling the hall's apparent height and reflecting the half 'sun' to form a whole.

Colin Martin

The Weather Project can be seen at Tate Modern in London until 21 March 2004. The work is the subject of an extensively illustrated book, Olafur Eliasson: The Weather Project, edited by the installation's curator, Susan May (Tate Publishing, £19.99).



J. ZIEHO/ELIASSON

the only route by which quantities of rare compounds such as eleutherobin or bryostatins can be made available for biological evaluation. Any synthesis, whether of a rare, highly potent natural product or of a more mundane yet equally important intermediate for a prescription drug, requires the careful selection of starting materials and potential synthetic routes.

Synthetic chemists already have a number of information sources, including online search sites such as *Crossfire Beilstein* (www.beilstein.com) and the American Chemical Society's *SciFinder* (www.cas.org). There are also printed reference works such as *Fieser's Reagents for Organic Synthesis* (Wiley), *Comprehensive Organic Synthesis* (Pergamon, now Elsevier) and of course *Organic Syntheses* (Wiley).

Organic Syntheses is the 'Mrs Beeton' for synthetic chemists. If you can find a 'recipe' for the reaction or compound you wish to make in these volumes, you can guarantee that the chemistry will work, because each entry is checked for accuracy, safety and

reliability. There is a cumulative index for the first 70 volumes, although many libraries do not have either this index or a complete set of 80 volumes. The release of the *Organic Syntheses Database* now provides access to all 80 on your computer screen.

So how does it perform? The user can search for structures and part-structures by importing these from one of their own drawing packages into a display box, and this is probably the most useful function. I tried several structures I wanted to synthesize and others that I knew to be within the volumes of *Organic Syntheses*, and obtained accurate links to the full text for all of them. Much less useful is the search for reaction types and named reactions. An enquiry about the aldol reaction elicited just one response, and no response was returned for queries on the Reformatsky and Heck reactions. Yet the term palladation produced two responses, both referring to syntheses submitted by Heck, and there are many others within the 80 volumes that have clearly been missed by this route.

Searches that use either keywords or author were generally very good, so searching for methodology associated with a key reagent, such as osmium tetroxide or palladium acetate, will elicit a large number of citations although, irritatingly, reaction names (Heck or Mitsunobu, for example) elicit nothing. The utility of the database is extended by the ability to search other Wiley databases, including EROS, the *Encyclopaedia of Reagents for Organic Synthesis* (although my login and password denied me access to explore these possibilities), and access to *Organic Reactions* will follow next spring.

Overall, the database is easy to use and will complement rather than replace the other existing online search tools. The one major advantage of the *Organic Syntheses Database* is that it provides access to an instant recipe that will work, rather than to a list of literature references in which the chemistry may or may not work.

John Mann is at the School of Chemistry, Queen's University Belfast, Belfast BT9 5AG, UK.

Tumour tactics

Hiroaki Kitano

Cancer is an extremely complex and heterogeneous disease that exhibits a high level of robustness against a range of therapeutic efforts. Robustness enables a system to maintain functionality in the face of various external and internal perturbations. Robustness is not an accidental property, but follows certain principles. It allows tumours to promote growth and survival in several ways. For example, heterogeneity among tumour cells provides a high level of redundancy, and hence increased chances of survival and growth; these benefits are further enhanced by feedback controls at the cellular level. Viewing cancer as a robust system may provide insight for the development of new drugs and therapies.

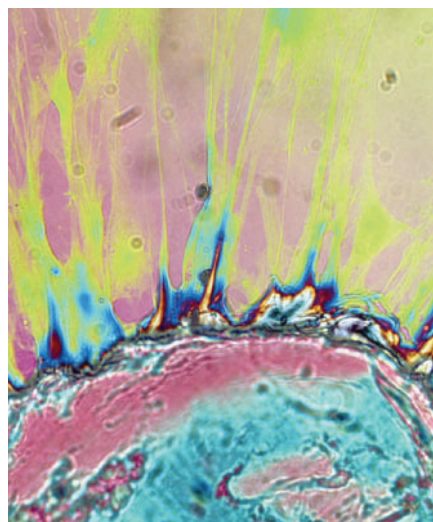
In general, carcinomas are heterogeneous and intrinsically redundant populations of cells, generated by transformations including diverse mutations in individual genes, mitotic recombinations, and the loss or gain of entire chromosomes (aneuploidy). This heterogeneous redundancy makes a tumour robust, because survival of any of these proliferating tumour cells allows the cancer to recur. Unfortunately, many anti-tumour drugs are mutagens, and so the diversity of mutations may be increased among surviving cells. In order to control this robustness, an effective therapy needs to reduce or at least avoid increasing this diversity.

This principle has a series of implications. As we are now beginning to understand the possible causes of heterogeneity of tumour cells, the obvious thing to do would be to avoid drugs that may increase heterogeneity by triggering further mutations or aneuploidy. The appropriate strategy will depend on the level of heterogeneity and on the drugs available. When the degree of heterogeneity is low, such as in early-stage, non-solid tumours, it would be best to avoid broad-spectrum cytotoxic drugs that may increase heterogeneity, and instead look for one that targets a specific molecule (for example, ABL protein for chronic myelogenous leukaemia). When the level of heterogeneity is high, there is a choice of two strategies. The first is to find a method that actively reduces diversity of mutation among tumour cells — possibly by eliminating all tumour cells except those with a certain mutation or cellular state — then using a drug that specifically targets cells with this property. The alternative would be to guide the malignant cells into dormancy and slow regression — rather than acute and dramatic regressions that impose selective pressures — so that heterogeneity does not increase. However,

the drugs used to induce dormancy must be carefully selected, as cytotoxic drugs might cause further mutations, resulting in an increase in heterogeneity despite dormancy, making resistance inevitable.

At the cellular level, feedback control enhances robustness against possible therapeutic efforts. This control, which protects normal cells by making them robust against perturbations, is an obstacle to tumour therapy. Tumour cells counter chemotherapy by overexpressing proteins, such as MDR1, that pump toxic chemicals out of the cell. This simple feedback regulation leads to the acquisition of multidrug resistance. Likewise, the protective role of p53 protein — which induces cells to undergo either apoptosis or cell-cycle arrest for DNA repair — is reduced by over-expression of mdm2, a suppressor of p53 protein. Thus, modulation of feedback control is a natural strategy for systems-level cancer treatments.

To implement such a strategy, we need to find a method that will systematically control the cell's activities, such as the cell cycle, growth decisions and apoptosis. Computer simulations have shown that a cell cycle that is robust against certain perturbations can be made extremely fragile when specific feedback loops are removed or attenuated, meaning that the cell cycle can be arrested with minimum perturbation. Although this theoretical result has not been experimentally verified, it suggests that robustness can be controlled by carefully selecting modulation targets and the systematic use of multiple drugs. This 'systems drug-discovery' approach aims to control the cell's dynamics, rather than its components, and although only speculative at present, may turn out to be critical for the



Hunt for fragility: weaknesses in tumour growth dynamics could yield new anti-cancer therapies.

Cancer robustness

Viewing cancer as a robust system with potential points of fragility opens up new strategies for the development of drugs and therapies.

development of new cancer therapies.

The ability to predict and control the cellular state may allow exploitation of the trade-off between robustness and fragility. Carlson and Doyle argue in their highly optimized tolerance (HOT) theory that robust complex systems show robustness against common perturbations and extreme fragility against unusual ones. HOT, which takes simple models from statistical physics and adds design or evolution aspects, was originally applied to technologically complex systems such as the Internet, but may be more widely applicable to biological and ecological systems, as demonstrated recently with models of forest fires and abstract lattice evolution. If this theory is relevant to tumours, the existence of extreme robustness implies that there will also be rare points of fragility which, if they can be found, would be a logical target for therapeutic efforts. Although finding such fragility might require a deep understanding of the tumour's regulatory network dynamics, there is potential for new system-level therapies. Ideally, it would be possible to exploit this trade-off to enhance robustness for a specific set of perturbations, so that fragility of the cancer system appears at the point where effective therapeutic agents are readily available.

Considering cancer as a robust system would provide us with a framework for future research strategies. Genetic heterogeneity of tumour cells may be an index of one aspect of robustness that can provide useful information on clinical strategies. An in-depth understanding of the dynamics of cellular robustness could enable us to find methods to control it. Such methods are most likely to be systematic low-dose use of multiple drugs. Future cancer therapies may be judged on their ability to help control the robustness of tumours.

Hiroaki Kitano is at Sony Computer Science Laboratories Inc., 3-14-13 Higashi-gotanda, Shinagawa, Tokyo 141-0022.

FURTHER READING

Kitano, H. *Nature* **420**, 206–210 (2002).
Morohashi, M. et al. *J. Theor. Biol.* **216**, 19–30 (2002).
Carlson, J. M. & Doyle, J. *Phys. Rev. Lett.* **84**, 2529–2532 (2000).

C. H. FOX/SPL

Essence of mitochondria

Katrin Henze and William Martin

For years, a unicellular creature called *Giardia* has occupied a special place in biology because it was thought to lack mitochondria. But it does have them — though tiny, they pack a surprising anaerobic punch.

All known life-forms are either prokaryotes or eukaryotes; there is nothing in between. Eukaryotic cells have a nucleus and compartments that are surrounded by two membranes, but prokaryotic cells never do. The prokaryotes came first; eukaryotes (all plants, animals, fungi and protists) evolved from them, and to this day biologists hotly debate how this transition took place, with about 20 different theories on the go¹. In most textbook accounts, an ancient prokaryotic lineage supposedly evolved a nucleus, giving rise to a eukaryote, which then acquired mitochondria — the double-membrane-bounded powerhouses of eukaryotic cells. A cornerstone of this view has been the single-celled eukaryote *Giardia intestinalis*, which many consider to be a primitive intermediate^{2,3}, a 'living fossil' from the time of the prokaryote-to-eukaryote transition, because it possesses a nucleus but lacks mitochondria. Or so we thought.

The paper on page 172 of this issue will surprise many people. There, Jorge Tovar and colleagues⁴ show that *Giardia* (Fig. 1) does have mitochondria after all. They are highly reduced, and are called mitosomes. Unlike the mitochondria familiar to most biologists, *Giardia*'s mitosomes do not generate ATP (energy). Rather, they are factories for the assembly of iron–sulphur (Fe–S) clusters, which *Giardia* requires to make ATP. These findings mark a turning point for views of early eukaryotic and mitochondrial evolution: *Giardia*'s place as an intermediate stage in standard schemes of eukaryotic evolutionary history^{2,3} is no longer tenable.

Giardia intestinalis (formerly *G. lamblia*) is a pathogen that shuns oxygenated environments and causes diarrhoea when it infects the human intestine. Like all cells, it requires a constant supply of ATP. Treatments for *Giardia* infection exploit the fact that its ATP-synthesizing biochemistry differs from ours. Whereas human cells make their ATP in mitochondria and require oxygen to do so, *Giardia* makes all of its ATP in the cytosol with the help of simple anaerobic pathways^{5,6}. Several enzymes of *Giardia*'s ATP-generating machinery contain Fe–S clusters^{5–9}, which are ubiquitous cofactors in the electron-transfer reactions involved in ATP production. They are essential for the parasite's survival. Among *Giardia*'s proteins that contain Fe–S clusters are hydrogenases^{7–9}, ferredoxins and

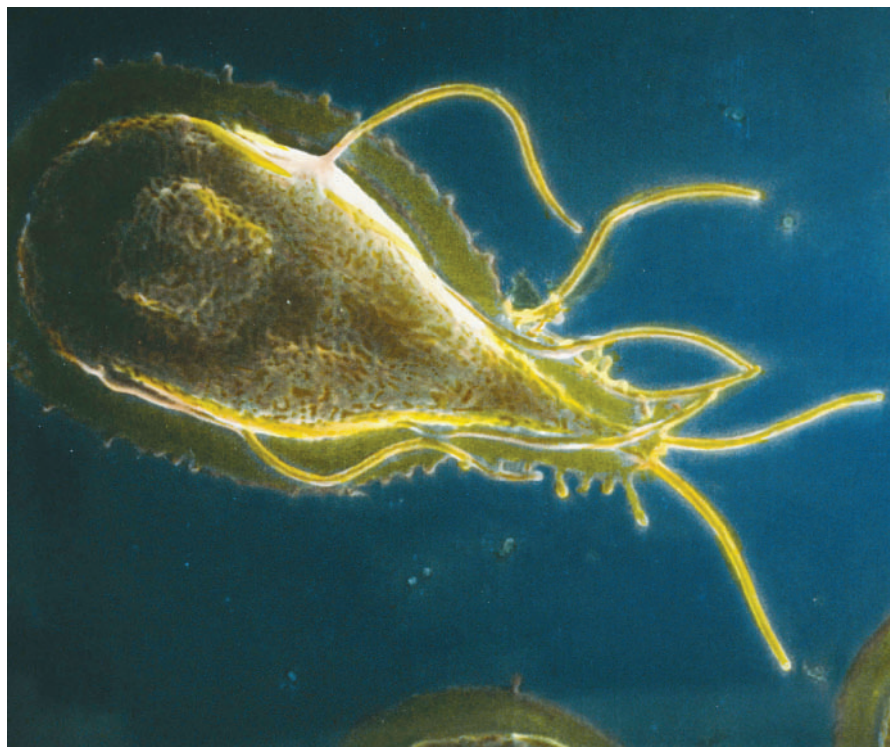


Figure 1 Mitochondria where least expected. *Giardia* has been the textbook example of a single-celled eukaryote that lacks mitochondria. Now, with the paper by Tovar and colleagues⁴, it becomes a prime example of an anaerobic eukaryote that possesses mitosomes — highly reduced mitochondria that do not function in core ATP synthesis, but are essential for the assembly of iron–sulphur clusters. The image is about 100,000 times life size.

pyruvate:ferredoxin oxidoreductase^{5,8,9} (PFO, which incidentally is the target for anti-*Giardia* drugs).

In other eukaryotes — yeast and humans, for example — the synthesis of Fe–S clusters always begins in mitochondria¹⁰. But if, as almost everyone believes, *Giardia* lacks mitochondria yet has Fe–S clusters, where are the clusters synthesized? Tovar *et al.*⁴ have investigated this question, and find the answer to be that they are made in curious little membrane-bounded structures, several dozen of which are present in each *Giardia* cell. These organelles are surrounded by two membranes, a decisive observation that leaves only two reasonable possibilities as to what the organelles might be. This is because, in a century of delving into cells, biologists have discovered only two kinds of organelle that are surrounded by a double membrane: chloroplasts and mitochondria.

Chloroplasts are the organelles of photosynthesis in plants, and are descendants of

cyanobacteria (photosynthetic prokaryotes) that, deep in evolutionary history, took up residence inside a eukaryotic host, forming a symbiotic partnership. But the *Giardia* organelles are almost certainly not chloroplasts, because no trace of chloroplast-specific biochemistry has ever been found in this protist. So are they mitochondria? Indeed they are.

Mitochondria are hallmarks of eukaryotic cells, and can be traced to another prokaryote — an α -proteobacterium¹¹ — that took up residence in its host long before chloroplasts arose. Mitochondria comprise a diverse family of organelles, including many anaerobic forms¹². Most notable among these are hydrogenosomes, hydrogen-producing forms of mitochondria discovered 30 years ago by Miklós Müller. As Tovar *et al.*⁴ explain, mitosomes are the most recent addition to the mitochondrial family of organelles. They are small, functionally reduced forms that were first found in the

HOSSLER, CUSTOM MED. STOCK PHOTO/SPL

amoeboid parasite *Entamoeba histolytica*¹³ and the pathogen *Trachipleistophora hominis*¹⁴ (which belongs to a group of fungi called microsporidia). Before those reports, *Entamoeba* and microsporidia were thought to lack mitochondria altogether. Now mitochondria appear in *Giardia*, too, and we have the first direct glimpse of their function.

The biochemical role of the mitosome can hardly be core ATP synthesis. That occurs in the cytosol in both *Entamoeba*⁵ and *Giardia*⁵. And it is uncertain that microsporidia even make their own ATP, because they can steal it from the cells that they infect¹⁵. Tovar *et al.*⁴ show that, instead, the *Giardia* mitosomes harbour critical enzymes of Fe–S cluster assembly. Furthermore, cell fractions enriched for the organelle assemble Fe–S clusters *in vitro*. This pins a function to the mitosome in *Giardia* and is a major advance in understanding.

Possibly, *Giardia* assembles Fe–S clusters in mitosomes for the same reason that yeast and humans assemble them in mitochondria: the process is very oxygen-sensitive. The mitochondrial matrix (the space inside the organelle) is the most oxygen-poor compartment in oxygen-respiring cells because oxygen is consumed in the surrounding membrane. Lloyd *et al.*¹⁶ have shown that *Giardia* possesses organelles that accumulate mitochondrion-specific dyes and that can transfer electrons from donors to acceptors, which could easily include oxygen. But it remains to be seen whether the organelles observed by Lloyd *et al.*¹⁶ and Tovar *et al.*⁴ are identical.

Although mitochondria are usually considered to be oxygen-dependent, *Giardia*'s tiny mitochondria have an anaerobic function (Fe–S cluster assembly) in the synthesis of oxygen-sensitive proteins such as hydrogenase and PFO, which are normally found in hydrogenosomes^{1,4–9,12}. Eukaryotes diversified while the oceans were largely anoxic^{17,18}, so these anaerobic functions are most easily seen as biochemical relicts of the mitochondrion's anaerobic past. Such anaerobic relicts are abundantly preserved in diverse eukaryotic lineages today^{9,12}.

We know that mitochondria arose as intracellular symbionts in the evolutionary past¹¹. But in what sort of host? That question still has biologists dumbfounded¹. In the most popular theories, *Giardia* is seen as a direct descendant of a hypothetical eukaryotic host lineage that existed before mitochondria did^{2,3}. But Tovar and colleagues' findings⁴ show that *Giardia* cannot have descended directly from such a host, because *Giardia* has mitosomes. So our understanding of the original mitochondrial host is not improved by these new findings, but our understanding of mitochondria certainly is. In its role as a living fossil from the time of prokaryote-to-eukaryote transition, *Giardia* is now retired. But it assumes a new place in

the textbooks as an exemplary eukaryote with tiny mitochondria that have a tenacious grip on an essential — and anaerobic — biochemical pathway.

Katrin Henze and William Martin are at the Institute of Botany, University of Düsseldorf, 40225 Düsseldorf, Germany.

e-mails: katrin.henze@uni-duesseldorf.de
w.martin@uni-duesseldorf.de

1. Martin, W., Hoffmeister, M., Rotte, C. & Henze, K. *Biol. Chem.* **382**, 1521–1539 (2001).
2. Sogin, M., Gunderson, J., Elwood, H., Alonso, R. & Peattie, D. *Science* **243**, 75–77 (1989).
3. Wheelis, M. L., Kandler, O. & Woese, C. R. *Proc. Natl Acad. Sci. USA* **89**, 2930–2934 (1992).
4. Tovar, J. *et al. Nature* **426**, 172–176 (2003).
5. Müller, M. in *Molecular Medical Parasitology* (ed. Marr, J.) 125–139 (Academic, London, 2003).
6. Lloyd, D. & Harris, J. C. *Trends Microbiol.* **10**, 122–127 (2002).

7. Lloyd, D., Ralphs, J. R. & Harris, J. C. *Microbiology* **148**, 727–733 (2002).
8. Cammack, R., Horner, D. S., van der Giezen, M., Kulda, J. & Lloyd, D. in *Biochemistry and Physiology of Anaerobic Bacteria* (eds Ljungdahl, L. G., Adams, M. W., Barton, L. L., Ferry, J. G. & Johnson, M.) 113–127 (Springer, New York, 2003).
9. Embley, T. M. *et al. IUBMB Life* **55**, 387–395 (2003).
10. Lill, R. & Kispal, G. *Trends Biochem. Sci.* **25**, 352–355 (2000).
11. Gray, M. W., Burger, G. & Lang, B. F. *Science* **283**, 1476–1481 (1999).
12. Tielens, A. G. M., Rotte, C., van Hellemond, J. & Martin, W. *Trends Biochem. Sci.* **27**, 564–572 (2002).
13. Tovar, J., Fischer, A. & Clark, C. G. *Mol. Microbiol.* **32**, 1013–1021 (1999).
14. Williams, B. A., Hirt, R. P., Lucocq, J. M. & Embley, T. M. *Nature* **418**, 865–869 (2002).
15. Katinka, M. D. *et al. Nature* **414**, 450–453 (2001).
16. Lloyd, D. *et al. Microbiology* **148**, 1349–1354 (2002).
17. Knoll, A. H. *Life on a Young Planet: The First Three Billion Years of Evolution on Earth* 122–160 (Princeton Univ. Press, 2003).
18. Theissen, U. *et al. Mol. Biol. Evol.* **20**, 1564–1574 (2003).

Chemistry

Dendrimers set to self-destruct

E. W. Meijer and M. H. P. van Genderen

The versatility of the branched macromolecules known as dendrimers is being exploited in various ways — explosively so, in the context of their application as potential drug-delivery systems.

If a good idea for scientific innovation emerges, you can be sure that several teams of researchers will be quickly on the case. An example comes in the form of three reports^{1–3} that explore the prospects of using dendrimers for drug release inside diseased cells.

Dendrimers are artificial macromolecules, constructed in step-by-step fashion using repetitive chemistry. The macromolecule constituents radiate in branching form from a central core, creating a sphere of chemical groups that can be tailored according to requirements. The results of the process are not only aesthetically appealing but offer chemists wonderful opportunities for exploring new ideas⁴. Dendrimers are large, but can be synthesized and characterized with a precision similar to that possible with smaller organic molecules. They do not suffer from the problem of 'polydispersity' that dogs linear macromolecules: that is, constituents of a given set of dendrimers have exactly the same molecular weight, rather than being a mixture of chains with a distribution of molecular weights. And the large number of identical chemical units in the branching units, as well as those at the periphery, confers great versatility. The end groups can be designed for various purposes, including sensing, catalysis or biochemical activity.

In this last instance, one of the potential virtues of dendrimers comes under the heading of 'multivalency': the enhanced effect that stems from lots of identical molecules being present at the same time and place. The combination of multivalency with precision architectures has made

dendrimers of increasing interest for biomedical applications, not least for drug delivery⁵. Dendrimers can enter cells remarkably easily, a property that means they have been investigated as potential gene-transfection agents. The chemical groups that bristle from the ends of the branches allow for tuning of biological properties, and can anchor one or more target groups onto the dendrimer. The compound that constitutes the drug itself can be physically encapsulated in the dendrimer or bound to it. There have been attempts to achieve total and simultaneous release of active agents through changing pH conditions. But generally the traditional route has been that of getting one chemical trigger to release one drug molecule.

Independently of one another, teams led by de Groot¹, Shabat² and McGrath³ have explored a much more advanced concept — simultaneous release of all of a dendrimer's functional groups by a single chemical trigger. All three exploit the fact that the dendrimer skeleton can be constructed in such a way that it can be made to disintegrate into known molecular fragments once the disintegration process has been initiated. Various terms have been used for these systems: "cascade-release dendrimers"¹, "dendrimer disassembly"³ and — colourfully — "self-immolative dendrimers"², these systems in effect perform a chemical amplification reaction. Triggered by a specific chemical signal, the dendrimer scaffold falls apart in several steps in a chain reaction, releasing all of the constituent molecules.

Two of the teams^{2,3} demonstrate the process in systems in which relatively simple

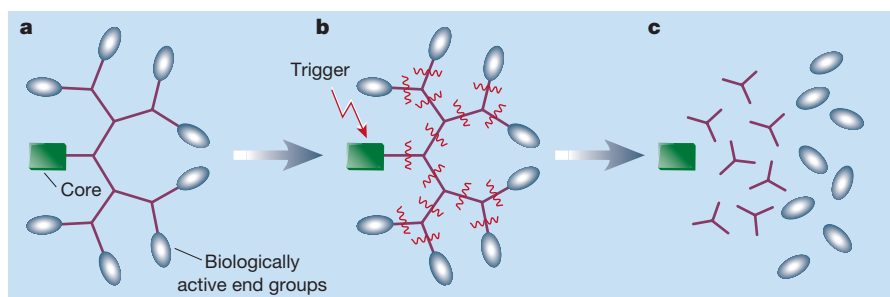


Figure 1 Bang on — a simplified depiction of the simultaneous release of biologically active end groups from a dendrimer^{1–3}. **a**, The basic dendrimer, shown here as a two-dimensional part of a sphere. **b**, Triggered by a specific signal, the dendrimer scaffold falls apart in a chain reaction. **c**, The result is release of all the constituent molecules, including the end groups. In the experiments of de Groot *et al.*¹, these end groups consisted of molecules of the anticancer drug paclitaxel (Taxol).

molecular fragments are released. De Groot and colleagues¹, however, have applied the principle in an especially elegant and appropriate way. They have not only devised methods for releasing the anticancer drug paclitaxel (Taxol), but also show that the dendrimer degradation products are not cytotoxic — except for paclitaxel itself, of course, which has the job of killing cancerous cells.

The first reaction activates the dendrimer core, initiating a cascade of 'elimination' reactions that lead to drug release (Fig. 1). Biodegradable polymers have been used before as drug carriers. But because dendrimers are so well defined, they allow fine control of the size, shape and composition of the release system. Their dendritic form, with many identical units, means that amplification can be achieved as a kind of explosion. A possible drawback, however, is the same that applies to every bomb — if the trigger is activated at the wrong time or place, the result will be devastating.

The concept described in the three

papers^{1–3} is intriguing, but will obviously need much more development before it can be put into practice in living cells. The next hurdle to overcome will be identifying a fuse that can be ignited to act as the trigger in biologically relevant conditions. Enzymatic reactions seem a promising avenue to explore. The pay-off of this approach, if it proves feasible, would be dendrimers that are specific for the enzymes present only in the cells to be targeted by a particular drug. ■

E. W. Meijer and M. H. P. van Genderen are in the Laboratory of Macromolecular and Organic Chemistry, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, the Netherlands. e-mail: e.w.meijer@tue.nl

1. de Groot, F. M. H., Albrecht, C., Koekkoek, R., Beusker, P. H. & Scheeren, H. W. *Angew. Chem. Int. Edn Engl.* **42**, 4490–4494 (2003).
2. Amir, R. J., Pessah, N., Shamis, M. & Shabat, D. *Angew. Chem. Int. Edn Engl.* **42**, 4494–4499 (2003).
3. Li, S., Szalai, M. L., Kevitch, R. M. & McGrath, D. V. *J. Am. Chem. Soc.* **125**, 10516–10517 (2003).
4. Bosman, A. W., Janssen, H. M. & Meijer, E. W. *Chem. Rev.* **99**, 1665–1688 (1999).
5. Patri, A. K., Majoros, I. J. & Baker, J. R. *Jr Curr. Opin. Chem. Biol.* **6**, 466–471 (2002).

Virology

Fresh assault on hepatitis C

Charles M. Rice

Hepatitis C virus causes severe liver disease. Initial trials of a newly developed agent that prevents the virus reproducing itself look promising. But what are the future prospects for this treatment?

As researchers gradually got to grips with viruses that attack the liver, two were identified early on — hepatitis A and hepatitis B. But clearly there were more, because there were also cases of hepatitis that had the hallmarks of neither virus. These cases occurred following blood transfusion, and for some time the resulting disease was known as non-A, non-B post-transfusion hepatitis. The culprit, hepatitis C virus (HCV), was identified in the late 1980s. Now, nearly 15 years later, the first small-molecule inhibitors of HCV are being tested in

humans. On page 186 of this issue¹, Lamarre *et al.* report spectacular success — at least in short-term trials — with an orally administered inhibitor of a viral protease, an enzyme that is essential for HCV to reproduce itself.

Hepatitis C is a severe medical problem. After the identity of the virus was revealed, the advent of reliable diagnostic methods led to the realization that here was a global infection afflicting some 170 million people. It is transmitted by contaminated blood, fortunately now rare in transfusions. Chronic —

persistent — infection is typical, and the disease course is unpredictable: some people seem to be unaffected, whereas others develop cirrhosis, end-stage liver disease and liver cancer. For many, a liver transplant is the only hope.

The virus itself has a tiny genome, consisting of a single-stranded RNA molecule only about 10,000 bases in length. This RNA serves not only as the template for viral replication, but as the viral messenger RNA for further viral production. It is translated into a long 'polyprotein', which is chopped up by cellular and viral proteases into at least 10 proteins that participate in RNA replication and new virus assembly.

One would think that this so-called simple virus would be easy to defeat. Not so. HCV has been remarkably adept at frustrating not just the human immune system, but also drug developers and virologists. It is not easy to study, in cell culture or in other animals. The virus replicates poorly in cell culture. Animal infection models are limited to the chimpanzee², or immunodeficient mice carrying engrafted human liver cells³. Only recently has it become possible to study how the RNA replicates^{4,5}, and how HCV enters cells^{6,7}, in cell culture.

The human immunodeficiency virus also has an RNA genome, which for obvious reasons has been the subject of intensive investigation. Following the success of protease inhibitors in suppressing HIV replication, the HCV-encoded serine protease emerged as a favourite target in the race for new anti-HCV drugs (the 'serine', incidentally, refers to a critical amino acid at the enzyme's active site). The enzyme itself resides towards one end, in the amino-terminal third, of a protein called NS3, and it cleaves the polyprotein at four downstream sites. No one knows precisely why, but the protease is essential for HCV replication⁸.

In early work, the protease was purified and its substrate specificity determined, and those in the field anxiously awaited determination of its structure. This was first reported^{9,10} in late 1996. But to the dismay of the drug designers and medicinal chemists who use structure as a guide, the substrate-binding channel of the protease was shallow and relatively featureless. It lacked the nooks, crannies and pockets that are used to select and refine highly specific and potent inhibitors. Many companies moved on, focusing on other HCV enzymes.

A few hardy groups pressed on, however, including Lamarre and his team. Key to the development of their successful antiviral agent, BILN 2061, was the observation that peptides mimicking amino-terminal cleavage products are competitive inhibitors^{11,12}. Starting with a weak enzyme inhibitor consisting of six amino acids — a hexapeptide — they crafted highly potent and specific hexapeptide inhibitors. They then trimmed the

size to a tripeptide to work towards creating a small molecule that could penetrate cells. Finally, they stabilized the compounds by intramolecular linking to create a molecule that had high potency in enzyme and cell-based assays. Tests with orally administered BILN 2061 in healthy volunteers showed that it seemed to have the right properties — lack of adverse effects and favourable retention and targeting in the body. So the team then launched short-term trials in HCV-infected patients.

As Lamarre *et al.* now report¹, the initial results of these trials are impressive. In patients given two doses of BILN 2061 per day for two days, virus levels declined by at least 100 to 1,000 times. This decrease in viral load is even more rapid than that observed for the current favoured therapy — administration of a modified form of interferon, pegylated interferon, by injection. Similar results were seen in patients who had had no treatment, those previously treated with interferon, and those with minimal or advanced liver disease.

It is still early in the fight against HCV. But the success of the first proof-of-principle trial of an HCV protease inhibitor is reason for optimism. The current treatment for HCV involves weekly injections of pegylated interferon, and daily oral administration of another drug, ribavirin. These drugs probably work nonspecifically by inhibiting HCV replication and stimulating the immune response. But they have numerous side effects. Moreover, the virus often clings on, and the most prevalent viral genotype in the United States and Europe, genotype 1, is eliminated in less than half of infected patients. More effective and well-tolerated drugs have been the dream of clinicians and their patients for many years. So is BILN 2061 the 'magic bullet'?

The next step will be trials to assess the longer-term safety and efficacy of BILN 2061. Experience with HIV suggests that a single antiviral drug is unlikely to eliminate HCV. RNA viruses have high mutation rates and survive by being flexible. On average, each RNA copy differs from the template RNA by about one mutation. A chronically infected person churns out around a trillion HCV particles each day. It is almost certain that inhibitor-resistant variants will be present and survive. Indeed, resistance has already been seen in cell culture studies¹³. Further, widely divergent isolates of HCV also exist around the globe, with six genotypes and numerous subtypes currently recognized. Only time will tell how a compound such as BILN 2061, optimized to attack one genotype, will fare against the others.

There is probably no magic bullet for HCV. Combination therapy, either with the existing treatments or with other specific antiviral drugs, will be needed to control this wily virus — a point that underscores

the importance of continuing to develop inhibitors against various viral targets. The work of Lamarre *et al.* provides a shining example that this is possible even against such a challenging target as the HCV protease. Let's hope that this news marks the beginning of a stampede of new compounds into clinical trials, and the dawn of new treatment options for hepatitis C.

Charles M. Rice is at the Center for the Study of Hepatitis C, Laboratory of Virology and Infectious Disease, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.

e-mail: ricec@rockefeller.edu

1. Lamarre, D. *et al.* *Nature* **426**, 186–189 (2003).

2. Lanford, R. E. & Bigger, C. *Virology* **293**, 1–9 (2002).
3. Mercer, D. F. *et al.* *Nature Med.* **7**, 927–933 (2001).
4. Lohmann, V. *et al.* *Science* **285**, 110–113 (1999).
5. Blight, K. J., Kolykhalov, A. A. & Rice, C. M. *Science* **290**, 1972–1974 (2000).
6. Bartosch, B., Dubuisson, J. & Cosset, F. L. *J. Exp. Med.* **197**, 633–642 (2003).
7. Hsu, M. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 7271–7276 (2003).
8. Kolykhalov, A. A., Mihalik, K., Feinstone, S. M. & Rice, C. M. *J. Virol.* **74**, 2046–2051 (2000).
9. Kim, J. L. *et al.* *Cell* **87**, 343–355 (1996).
10. Love, R. A. *et al.* *Cell* **87**, 331–342 (1996).
11. Llinàs-Brunet, M. *et al.* *Bioorg. Med. Chem. Lett.* **8**, 1713–1718 (1998).
12. Steinkühler, C. *et al.* *Biochemistry* **37**, 8899–8905 (1998).
13. Trozzi, C. *et al.* *J. Virol.* **77**, 3669–3679 (2003).

This article was published online on 26 October 2003
(doi:10.1038/nature02105).

Materials science

Close-up on cracks

Jay Fineberg

How do things break? The fracture of materials is part of our everyday experience, and yet the process is not well understood. A study of crack propagation at microscopic scales shows the devil in the details.

The effects of material failure are obvious on macroscopic scales, and yet the dynamics of fracture that lead to such failure are governed entirely by the material's behaviour at the smallest scales. This fact is underscored by work presented by Buehler *et al.*¹ on page 141 of this issue. Ninety years after Inglis pointed out that cracks in a material play a key role as stress enhancers, Buehler *et al.* show that the dynamics of a crack change unexpectedly as the huge (putatively infinite) stresses generated around the crack's tip alter the material properties across a microscopic region.

Once a crack starts to develop in a material, what determines its dynamics? You might think that the dynamics of fracture are self-evident: huge stresses at the tip of a crack cause molecular bonds to rupture, and the crack lengthens ever more rapidly. But that is only the beginning of the story of how different materials cope with the mathematical singularity at the tip of a crack that, as a crack becomes infinitely sharp, gives rise to infinite stresses.

The theoretical framework for the formation of the stress singularity is based on the assumption that a material under stress follows the rules of linear elasticity — that, like a common spring, the amount that a material stretches is proportional to how hard it is pulled. In nearly all cases, this assumption is perfectly valid if a material is only slightly deformed from its initial state. Fracture, on the other hand, is a nonlinear process and the model of spring-like behaviour breaks down around the tip of a crack when the stresses approach the mathematical singularity predicted by Inglis. The way that a particular

material acts to 'blunt' this singularity ultimately determines its strength, its mode of failure and the dynamics that describe it. Both the rapid failure that is typical of low-strength, brittle materials (such as glass) and the gradual failure of high-strength, ductile materials (such as steel) are governed by the microscopic response of the material around crack tips.

There have been numerous attempts^{2–6} to model this near-tip zone, but progress has been hampered by lack of knowledge of the detailed physical processes at these quasi-infinite stress levels. Experimental access to this region is extremely difficult — the zone is small, sometimes approaching atomic dimensions, and fracture can occur very rapidly (brittle fracture can happen at speeds approaching the speed of sound in the material). Computational advances^{7–9} now offer a way in.

Buehler *et al.*¹ ran simulations of millions of interacting atoms to analyse the fine details of the fracture process in two qualitatively different types of (artificial) material. These materials are conceptually simple, but their properties are designed to elucidate the very real effects of the nonlinear material response to the large stresses present around a crack tip. The force between neighbouring atoms in both materials is allowed to behave in a linear (or elastic) way until their separation exceeds a critical value. Beyond that value, a 'hyperelastic' force law applies: in one material, the effective spring constant between the neighbouring atoms is doubled (called hyperelastic stiffening); in the other material it is halved (hyperelastic softening). Any pair of atoms separated by more than a second



100 YEARS AGO

It is reported in some of the daily papers that Dr. Otto Schmidt, of Cologne, has succeeded in isolating and cultivating a parasite from cancer and in preparing an antiserum for the disease. So many positive statements of the isolation of a cancer-parasite have been made during the last few years, and have subsequently proved to be incorrect, and so many capable men have been investigating cancer without result, that reports of this kind cannot be accepted without further proof. The publicity given to matters of this kind is much to be deprecated; in the majority of instances false hopes are raised which must end in disappointment for many sufferers.

From *Nature* 12 November 1903.

50 YEARS AGO

It was a famous moment in the history of science when, during the discussion of Darwin's theory of evolution at the British Association meeting at Oxford in 1860, Bishop Wilberforce turned to T. H. Huxley and asked him whether he claimed descent from an ape on his father's or his mother's side. The actual words of Huxley's reply are not known... The main source of our information, his son Leonard Huxley, wrote "most unluckily, no contemporary account of his own exists of the encounter". Such an account does, however, exist in a letter written to Dr. Dyster within a few months of the meeting, on September 9, 1860, and now preserved in the collection of Huxley Papers at the Imperial College of Science and Technology, London... "When I got up I spoke pretty much to the effect — that I had listened with great attention to the Lord Bishop's speech but had been unable to discover either a new fact or a new argument in it — except indeed the question raised as to my personal predilections in the matter of ancestry... If then, said I, the question is put to me would I rather have a miserable ape for a grandfather or a man highly endowed by nature and possessing great means and influence and yet who employs those faculties and that influence for the mere purpose of introducing ridicule into a grave scientific discussion — I unhesitatingly affirm my preference for the ape. Whereupon there was unextinguishable laughter among the people, and they listened to the rest of my argument with the greatest attention."

From *Nature* 14 November 1953.

critical value are allowed to detach, thereby enabling fracture to occur.

The calculations show how significant the effects of hyperelasticity can be, even when the hyperelastic zone is only a few hundred atoms in size and the remainder of the material is behaving elastically. Buehler *et al.*¹ suggest that, under the right conditions, the properties of this tiny hyperelastic region entirely control the fracture process. Supersonic propagation of cracks — far surpassing the limiting velocity allowed by linear elastic theory¹⁰ — becomes possible, and as a result shock fronts may be emitted from the hyperelastic region.

As every crack is surrounded by a non-linear region, hyperelastic effects should be commonplace. On the other hand, laboratory experiments¹¹ have indicated that linear elastic theory provides an excellent description of the dynamics of rapidly moving cracks. So when would the more exotic behaviour induced by hyperelastic effects be seen? Buehler *et al.*¹ predict that the key to this question is the ratio of the size of the hyperelastic region marking the onset of hyperelasticity, to an 'energy' length scale, defined as the size of the region around the tip that encompasses enough energy to drive the fracture process. When the hyperelastic region is much smaller than the energy length scale, hyperelastic effects are negligible; but if the two are similar, these effects could dominate the fracture process. We might then expect to observe the effects of hyperelastic behaviour when fracture occurs in a highly strained material, or if a material suffers a high rate of strain, such as caused by the impact of a projectile.

The process of rapid fracture is strongly influenced by the interplay of physical effects on many different scales. Simulations, such as these by Buehler *et al.*¹, should enable us to bridge the gap between the laboratory scales where fracture is observed and the near-atomic scale where fracture germinates. Describing the detailed motion of millions of atoms involved in the fracture process is an impressive technical feat, but it is still important to be able to see the wood for the trees — to identify the key features relating the delicate interplay between the myriad dancing atoms at a crack tip and the macroscopic effects that they generate. This work, together with other studies^{7–9}, is a step towards both a fundamental understanding of these processes and, possibly, a powerful tool for the design of new materials. ■

Jay Fineberg is at the Racah Institute of Physics, Hebrew University of Jerusalem, Givat Ram 91904, Jerusalem, Israel.

e-mail: jay@vms.huji.ac.il

1. Buehler, M. J., Abraham, F. F. & Gao, H. *Nature* **426**, 141–146 (2003).
2. Langer, J. S. & Lobkovsky, A. E. *J. Mech. Phys. Solids* **46**, 1521–1556 (1998).
3. Fleck, N. A. & Hutchinson, J. W. *J. Mech. Phys. Solids* **49**, 2245–2271 (2001).
4. Aranson, I. S., Kalatsky, V. A. & Vinokur, V. M. *Phys. Rev. Lett.* **85**, 118–121 (2000).
5. Karma, A., Kessler, D. A. & Levine, H. *Phys. Rev. Lett.* **87**, 045501 (2001).
6. Needleman, A. & Tvergaard, V. *Int. J. Fracture* **101**, 73–97 (2000).
7. Abraham, F. F., Brodbeck, D., Rafee, R. A. & Rudge, W. E. *Phys. Rev. Lett.* **73**, 272–275 (1994).
8. Holland, D. & Marder, M. *Phys. Rev. Lett.* **80**, 746–749 (1998).
9. Gumbsch, P., Zhou, S. J. & Holian, B. L. *Phys. Rev. B* **55**, 3445–3455 (1997).
10. Freund, L. B. *Dynamic Fracture Mechanics* (Cambridge Univ. Press, 1990).
11. Sharon, E. & Fineberg, J. *Nature* **397**, 333–335 (1999).

Plant development

An axis of auxin

Stefan Kepinski and Ottoline Leyser

Embryos have two distinct ends, which become apparent early on. Quite how this initial polarity is sustained in plant embryos has been unclear. Step forward the *agent provocateur* of plant development — auxin.

In multicellular organisms, different kinds of cell are specialized for different tasks — reproduction, say, or light perception. Clearly, the proper functioning of these organisms requires that the various cell types are positioned correctly relative to one another. Hence, common to the development of all multicellular organisms is the formation of axes along which the body plan is organized. Such axes are usually manifest very early in development, and indeed it is often impossible to identify an apolar stage¹. This is certainly the case in many higher plants, including the experimental organism *Arabidopsis thaliana*, where a developmental decision that ultimately gives rise to a shoot

at one end and a root at the other can be traced back to the first, asymmetric division of the fertilized egg cell². But how is this initial asymmetry maintained and translated into the polar root–shoot axis? On page 147 of this issue³, Friml and co-workers strikingly demonstrate that such 'elaboration' depends on the movement to and fro of the plant hormone auxin.

As Fig. 1 shows, the initial division of the fertilized *Arabidopsis* egg cell produces two daughter cells: a small upper cell (the apical cell) and a larger basal cell. The apical cell generates the 'proembryo', which develops through a series of stereotypic divisions to give rise to the upper, central and mid-lower

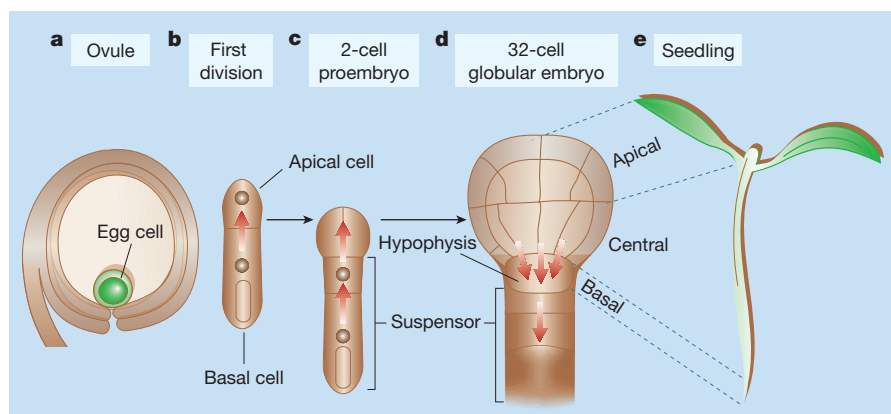


Figure 1 The basics of early *Arabidopsis* development. **a**, The egg cell lies deep in the female parts of the flower, within the ovule. **b**, Once fertilized, the egg undergoes an asymmetric division, producing a larger basal cell and a smaller apical cell. Friml *et al.*³ have found that the basal cell exports the hormone auxin (red arrow) and that the apical cell responds to it. **c**, The basal cell divides horizontally to form the suspensor, while the auxin-receiving apical cell divides vertically to form the 2-cell proembryo. **d**, Around the stage of the 32-cell globular embryo, the flow of auxin is suddenly reversed³ and now accumulates in the hypophysis, the founder of the basal regions of the embryonic root. At this stage, one can begin to distinguish the regions that give rise to the apical, central and basal parts of the seedling (**e**).

regions of the seedling. Meanwhile, the basal cell produces the suspensor — a stack of cells that attaches the proembryo to maternal tissue. Initially, all of the suspensor cells lie outside the embryo proper, but later the uppermost cell is recruited by the proembryo to form the hypophysis, the founder of the lowest regions of the embryonic root².

It has been difficult to pin down how the initial asymmetry is translated into this suspensor–root–shoot axis. One contender for the role is auxin. A truly multi-talented signalling molecule, auxin has a finger in virtually every plant-developmental pie, from the control of branching in both shoot and root to the patterning of the root tip⁴. All of these processes depend on the cellular responses to local auxin concentrations, and on the generation of patterns of auxin accumulation. The responses to auxin involve changes in gene expression, which are mediated by families of positive and negative gene regulators whose relative abundance is tightly controlled by auxin⁵. The accumulation of auxin, meanwhile, is directed by two protein families: the AUX influx carriers, which pass auxin into cells, and the PIN efflux carriers, which pass it out⁶. Directionality of auxin transport is provided by the asymmetric localization of the PIN proteins⁶.

A role for auxin in elaborating the embryonic axis has been suspected for many years, not least because polarity defects can be induced in embryos by blocking auxin movement⁷. More recent molecular genetic evidence comes from the analysis of mutations in three *Arabidopsis* genes, *MONOPTEROUS* (*MP*), *BODENLOS* (*BDL*) and *GNOM* (*GN*), all of which cause defects in axis elaboration. *MP* encodes a positive and *BDL* a negative regulator of auxin-

inducible genes^{8,9}, whereas *GN* encodes a protein needed for the proper subcellular targeting of the PIN proteins^{10,11}. What is exciting about the work of Friml *et al.*³ is that it ties these strands together definitively.

To follow the responses of embryos to auxin, the authors introduced a gene for green fluorescent protein (GFP), attached to a synthetic control region that is activated by auxin. They found that the first signs of a response (that is, the first signs of GFP production and hence auxin activity) occur at a very early stage, in the apical daughter cell of the asymmetrically divided egg. As the proembryo develops, the auxin-response signal persists, remaining absent from the suspensor cells beneath. This pattern continues until the proembryo consists of about 32 cells, when a remarkable thing happens. The axis of auxin response is suddenly reversed, becoming undetectable in the apical regions, with a new maximum in the developing hypophysis beneath. The authors find that these response patterns reflect actual gradients of auxin concentration, with maximal responses occurring at maximal concentrations.

To find out how these gradients arise, Friml *et al.* investigated the expression patterns of the PIN auxin-efflux proteins. They show that, within the two-cell proembryo, a previously uncharacterized PIN-family member — PIN7 — is expressed in the basal cell at the boundary facing the cell's apical sister. This is consistent with the auxin maximum in the apical cell. Later, the cells of the suspensor continue to express PIN7 at their apical side, while in the proembryo another protein, PIN1, is expressed without apparent polarity¹⁰.

But at the 32-cell stage, it's all change. Both PIN1 and PIN7 become localized to the

basal membranes of the cells in which they are expressed (proembryo and suspensor cells, respectively) — an event that coincides with the reversal of the auxin gradient and, presumably, with the onset of auxin production in the proembryo. This new direction of auxin flow is apparently reinforced by the expression of two other PIN-family members, so that, although PIN7 is positioned to transport auxin down the suspensor and out of the embryo, the net effect is the accumulation of auxin, and a maximal response to it, in the embryonic root.

Do these auxin fluxes help to maintain the embryonic axis of polarity? The authors' studies of mutant plants show that they do. Embryos with mutations in PIN7 have trouble establishing the initial auxin-response maximum in the apical cell and its daughters, and this coincides with a confused apical/basal identity in the proembryo. The effects of mutations in other PIN proteins are milder and affect later stages of basal embryo development. But matters are not entirely predictable. Knowing the PINs' expression patterns, if you were going to put money on any of these mutants not making it out of embryogenesis it would be on those with PIN7 defects — yet these plants recuperate to produce relatively normal seedlings. Interestingly, the recovered axis is always in the correct orientation, hinting that the polarizing influence of the suspensor is maintained into these later stages. It is not clear how polarity is restored, but it does require auxin efflux, as plants with mutations in all four embryonically expressed PINs fail to recover. These data, together with the requirement for an embryonic response to auxin implicit in the defects caused by *MP* and *BDL* mutations, underline the relevance of asymmetric auxin transport and responses in maintaining polarity.

So does auxin do it all? Is the initial polarizing signal from maternal tissue — the signal that directs the first, asymmetric cell division — also auxin? And can auxin itself direct the polar localization of PIN proteins? It is certainly possible that the initial asymmetry is directed by a low level of auxin flow from maternal tissues that passes beneath the radar of current techniques, and that maternal auxin, channelled up the suspensor, could continue to provide axial information later on. Furthermore, auxin has for many years been proposed to act in feedback loops to regulate its own flux⁶. But it is harder to envisage a mechanism whereby auxin could trigger the dynamic changes in its direction of flow that are observed here.

Are there then other signals, and, if so, what is their relationship to auxin? Is auxin instructive or permissive? If the former, auxin must first instruct apical, and then, almost immediately, basal fates. As this would require that auxin concentration is

maintained and interpreted with the utmost rigour, auxin might instead permit development, within an apical or basal context set by additional input. At least we are now in a position to muse on these possibilities, thanks to the work of Friml and colleagues. ■

Stefan Kepinski and Ottoline Leyser are in the Department of Biology, University of York, York YO10 5YW, UK.
e-mails: ssk2@york.ac.uk
hml1@york.ac.uk

Condensed-matter physics

The quest for imperfection

Thomas F. Rosenbaum

The electrical properties of silver chalcogenides are unusually affected by magnetic fields. A simulation suggests how this might arise from tiny imperfections and could facilitate the design of new materials.

The change in the electrical resistance of a material with magnetic field — its magnetoresistance — lies at the heart of many applications, from storing information in a computer to measuring the speed of a car. The resistance of two materials in particular, silver selenide and silver telluride, can be made to increase linearly as the applied magnetic field increases, up to field values at least one million times that of the Earth's magnetic field^{1,2}. This linear response over such a large range makes these semiconductor materials attractive as magnetic-field sensors, although their performance is unusual in terms of basic physics. Parish and Littlewood³ have tackled the question, as they report on page 162 of this issue, using a combination of fundamental theory and computer simulations. They show that it is not material perfection, but macroscopic disorder and inhomogeneity, that are the essential ingredients for such a remarkable magnetoresistive response.

Crystalline perfection and purity is the paradigm for the semiconductor electronics industry, where gowned technicians in dust-free clean rooms prepare crystal wafers of silicon containing impurities controlled to one part in a billion. In contrast, perfectly stoichiometric, homogeneous silver selenide and silver telluride (Ag₂Se and Ag₂Te, respectively), and 'silver chalcogenides' collectively are useless for applications. It is only when a small amount of excess silver is introduced, ideally at the level of one part in 10,000, that the material becomes interesting. The exact distribution of the excess silver atoms in the 'optimally doped' material is not known, but the atoms presumably group in small clusters or fine wires on nanometre or micrometre scales.

So how do such fluctuations in composition lead to an anomalously large, linear magnetoresistance that also doesn't saturate

1. Wolpert, L. *Principles of Development* (Oxford Univ. Press, 2002).
2. Jürgens, G. *EMBO J.* **20**, 3609–3616 (2001).
3. Friml, J. *et al. Nature* **426**, 147–153 (2003).
4. Berleth, T. & Sachs, T. *Curr. Opin. Plant Biol.* **4**, 57–62 (2001).
5. Kepinski, S. & Leyser, O. *Plant Cell* **14**, S81–S95 (2002).
6. Friml, J. *Curr. Opin. Plant Biol.* **6**, 7–12 (2003).
7. Hadfi, K., Speth, V. & Neuhaus, G. *Development* **125**, 879–887 (1998).
8. Hardtke, C. S. & Berleth, T. *EMBO J.* **17**, 1405–1411 (1998).
9. Hamann, T. *et al. Genes Dev.* **16**, 1610–1615 (2002).
10. Steinmann, T. *et al. Science* **286**, 316–318 (1999).
11. Geldner, N. *et al. Cell* **112**, 219–230 (2003).

at high field values? Abrikosov⁴ was the first to realize that the problem should be modelled in terms of the creation of an 'effective medium', in which disorder is combined with quantum mechanics in a theory that explains the properties of semiconductors that are almost, but not quite, metals. Parish and Littlewood³, however, invoke a classical explanation that is applicable over a broad range of temperatures and to a wide variety of compositional irregularities. They construct (conceptually, and on the computer) a two-dimensional, random network of resistors — effectively, an inhomogeneous semiconductor. Usually, a resistor has a single current input and output, but these resistors are special: each disk-like resistor has four terminals around its edge, two for the input and output of current and two additional posts for reading voltages. In this way, the voltage drop across the resistor need not be aligned with the direction of the current. From such a four-terminal device it is possible to construct a matrix of the device's conductivity: diagonal components correspond to the voltage being aligned with the current; off-diagonal components arise because a current in one direction can induce a voltage in a different direction.

There are two key elements to consider — the effect of an applied magnetic field, and the response of the resistor network as a whole. When a magnetic field is applied perpendicular to the direction of current flow, the charge-carrying electrons in each disk are forced into circular trajectories (called cyclotron orbits) and a voltage is induced that is transverse to both the field and the current directions. This transverse voltage is known as the Hall voltage and has the special property of growing linearly with increasing magnetic field. This linear response is obviously necessary for matching theory to experiment, but it is not enough to explain

the exact behaviour of the silver chalcogenides. The magnetoresistance would still be expected to saturate at a magnetic-field strength that is almost 100 times lower than that already reached in experiments — experiments that show no evidence of saturation.

It is at this point that the collective behaviour of a random array of resistance disks becomes important. No longer do the cyclotron orbits alone set the definitive scale of the response. Rather, disorder and inhomogeneity, as expressed by fluctuations in the properties of the individual array resistors — for example, whether the Hall voltage is positive or negative — are the appropriate control parameters. The current paths no longer simply follow electric-field lines, but can loop back on themselves and even travel perpendicular to the applied voltage. The net effect in a large array is a linear magnetoresistance that corresponds to that of the silver chalcogenides. Moreover, Parish and Littlewood's approach in essence presents design rules for creating new materials with a tailor-made magnetoresistive response.

The concept of exploiting imperfection can be applied across the field of condensed-matter physics. Heterogeneous systems, from polymers to high-temperature superconductors, organize themselves on microscopic length scales with macroscopic consequences. Eschewing the rigorous demands and expense of perfectly replicated and ordered devices has obvious advantages, but introduces complications in addressability and control. A good example is the nascent field of quantum computation. Here it is feared that disorder will lead to dissipation, disrupting the finely tuned superposition of quantum states that is necessary for processing quantum information. Yet quantum entanglement can, in fact, dominate the magnetic response of a solid-state system⁵.

Rather than building gates between components in a conventional sense, it may be possible to take advantage of materials that have different microscopic environments, and which are tuned and addressable electronically, magnetically or optically. Following this strategy, disorder will be sought for its natural mix of physical parameters and length scales, permitting the development of flexible and easily scalable devices. ■

Thomas F. Rosenbaum is in the Department of Physics and the James Franck Institute, University of Chicago, 5640 Ellis Avenue, Chicago, Illinois 60637, USA.
e-mail: t-rosenbaum@uchicago.edu

1. Xu, R. *et al. Nature* **390**, 57–60 (1997).
2. Husmann, A. *et al. Nature* **417**, 421–424 (2002).
3. Parish, M. M. & Littlewood, P. B. *Nature* **426**, 162–165 (2003).
4. Abrikosov, A. *Phys. Rev. B* **58**, 2788–2794 (1998).
5. Ghosh, S., Rosenbaum, T. F., Aeppli, G. & Coppersmith, S. N. *Nature* **425**, 48–51 (2003).

Quantum physics

A light touch on quantum strangeness

Phys. Rev. Lett. **91**, 180402 (2003)

Telecommunications engineers don't normally worry about the fine details of quantum theory. Encoding information in light pulses, sending the pulses through fibre-optic networks and routing them with multiplexers and amplifiers: it all looks relatively routine. But Nicolas Brunner and co-workers say that optical telecommunications engineers are unwittingly working at the forefront of quantum information theory. Standard processes in fibre-optic networks embody fundamental features of quantum logic, they claim, and could provide engineers with a simple quantum-based formalism for understanding what happens to their light signals.

At the core of the matter are processes that depend on the polarization of light pulses. Different polarization modes travel at different speeds in optical fibres. This is equivalent to making a 'weak' measurement on a quantum system (in this case, the photons of the pulse), a measurement that minimizes the 'back action', the process through which a quantum wavefunction collapses. And the polarization-dependent signal losses that occur in optical devices such as amplifiers amount to a 'postselection' of polarization states; in essence, forcing the wavefunction to collapse into a particular state. These processes are often seen as esoteric aspects of quantum theory. Brunner *et al.*, however, point out that they happen all the time in optical networks.

Philip Ball

Physiology

Fast food

Oikos **103**, 397–402 (2003)

Wading birds are record-breaking eaters, claim Anders Kvist and Åke Lindström. They find that, before migration, the birds can take on fuel at rates beyond those previously documented in vertebrates.

Kvist and Lindström caught individuals of 15 Swedish wader species in the autumn, just before they left for warmer climes, and provided them with an all-you-can-eat mealworm buffet. Most species could assimilate energy at more than six times their basal metabolic rate, which is unusually high for vertebrates. But the champion glutton was the red knot, *Calidris canutus* (pictured), which assimilated energy at more than 10 times its basal metabolic rate — the previous record was 7.7, in the mouse. This level of consumption allowed the birds to gain weight at almost 15% per day.

Red knot use their reserves to power



Red knot — here seen eating out.

an epic migration: they spend several months in transit each year, flying from the Arctic to deep in the Southern Hemisphere, and then back again. Kvist and Lindström conclude that the upper limit for energy assimilation in vertebrates has probably been underestimated, and that the limit in a particular species depends on the digestive system and the capacity to lay down fat.

John Whitfield

Hydrology

White Nile past

Geology **31**, 1001–1004 (2003)

The secret history of the White Nile — one of the two main tributaries of North Africa's greatest river — has been revealed using a combination of high-resolution satellite imagery and sediment-dating techniques. Martin A. J. Williams and colleagues find that a large lake, now disappeared, formed as much as 400,000 years ago and explains the tributary's unusually gentle gradient.

The White and Blue Nile feed into the Nile itself at Khartoum. The White Nile's hydrological history is known from about 15,000 years ago. But Williams *et al.* have gone back further in time by identifying the shoreline of an ancient lake from satellite images, and dating quartz grains in the relevant sediments by a method known as optically stimulated luminescence.

The result is a series of snapshots of the river's behaviour over time — which, most notably, includes a giant but quite shallow lake some 500 km long in a north–south direction, reaching almost to present-day Khartoum, and up to 70 km wide. The authors point out that there has been more than a century of speculation that such a feature once existed, to account for the White Nile's unusual flatness. The mean longitudinal gradient is a less than dizzying 1 in 75,000.

Tom Clarke

Virology

Suppressing a terminator

Cell **115**, 319–331 (2003)

A protein called reverse transcriptase plays a crucial part in the life of retroviruses — viruses whose genomes consist of RNA rather than DNA. This enzyme (which is

encoded by the viral *pol* gene) copies RNA genomes into duplex DNA, which then integrates into the DNA of the host cell, allowing more viral RNAs to be made. Marianna Orlova *et al.* now reveal another role for reverse transcriptase — and another way in which viruses can subvert host processes.

When organisms make a protein, they first copy the encoding gene into messenger RNA; that mRNA is then translated into protein. In murine leukaemia virus, a single mRNA is used to express two genes, *gag* and *pol*. But 'stop' signals at the end of *gag* would normally halt translation of this multifunctional mRNA before *pol* is reached. How can reverse transcriptase be produced?

Translation termination involves several proteins; one, eRF1, recognizes stop sequences. Orlova *et al.* now show that existing reverse transcriptase binds to eRF1. This interaction does not affect the enzyme's ability to make DNA. But it does reduce the amount of free eRF1, such that the stop signal cannot be recognized — allowing more *gag-pol* mRNA to be translated. If the interaction is blocked, the levels of *gag-pol* translated are not enough to support viral propagation.

Angela K. Eggleston

Psychophysics

Who are you?

Proc. Natl Acad. Sci. USA **100**, 13105–13110 (2003)

The brain centres involved in recognizing faces and in processing facial expressions have long been thought of as entirely separate — partly because brain-damaged patients with 'prosopagnosia' cannot recognize faces (even their own), but can still distinguish expressions. Beatrice de Gelder *et al.* show that the two processes may, in fact, be linked.

The authors found that prosopagnosic patients were better able to recognize which face part belonged to a given face if the faces were happy or angry rather than neutral. By contrast, control subjects found the task more difficult when the faces showed emotion. Using functional magnetic resonance imaging, the team also found that, in the prosopagnosics with damage to the two typical face-recognition areas in the brain, the task activated other areas. These included the amygdala, which processes emotion, and the superior temporal sulcus, which helps deal with social interactions.

The authors conclude that activation of these areas boosts face recognition by prosopagnosics, although the exact mechanism involved is unknown. The finding also suggests that diagnosis of face-recognition disorders — and possibly the treatment of prosopagnosia — should exploit the interactions between emotional skills and impaired recognition of identity.

Helen Pearson

Radar imaging of the lunar poles

Long-wavelength measurements reveal a paucity of ice in the Moon's polar craters.

We have used a radio telescope at Arecibo Observatory, Puerto Rico, to map features of the lunar poles — some as small as 300 metres across — by collecting long-wavelength radar images that can penetrate several metres of lunar dust. We find that areas of the crater floors at the poles that are in permanent shadow from the Sun, which are potential cold traps for water or other volatiles, do not give rise to strong radar echoes like those associated with thick ice deposits in the polar craters on Mercury. Any lunar ice present within regions visible to the Arecibo radar must therefore be in the form of distributed grains or thin layers.

Areas of the lunar polar regions are in permanent shadow with respect to Solar illumination^{1,2}. The presence of significant quantities of water ice in these regions has been inferred from bistatic radar data from the Clementine mission^{3,4} and from neutron-spectrometer measurements made by the Lunar Prospector⁵. Radar probing can be used to study the physical distribution of ice deposits: where ice layers are thicker than several times the illuminating wavelength and are characterized by density inhomogeneities (internal cracks or suspended rocks), a phenomenon called 'coherent backscatter' can produce strong radar returns and a distinctive polarization signature⁶. This type of echo, at wavelengths of 3.5–70 cm, is observed for many permanently shadowed crater floors near both poles of Mercury, and has been interpreted as indicating the presence of thick ice layers^{7,8}.

Earth-based radar measurements collected previously at wavelengths of 3.5 cm (ref. 1) and 12.6 cm (ref. 9) did not reveal strong backscatter (or the distinctive polarization signature of ice) from lunar polar terrain in permanent shadow, with the exception of very rugged crater ejecta and some radar-facing crater walls. If these earlier radar observations did not probe the lunar regolith to sufficient depth, however, there could still be thick ice layers tens of centimetres below the surface. Our radar measurements, taken at a wavelength of 70 cm, can penetrate several metres of lunar dust.

We used the 430-megahertz radar system at Arecibo Observatory to acquire individual delay-Doppler 'looks' with a surface spatial resolution of about 300 m. A circularly polarized signal was transmitted, and we received the opposite-sense circular return. The radar looks were converted to a selenographic coordinate format, normalized to their relative noise levels, and summed to form a final map. Lunar libration conditions

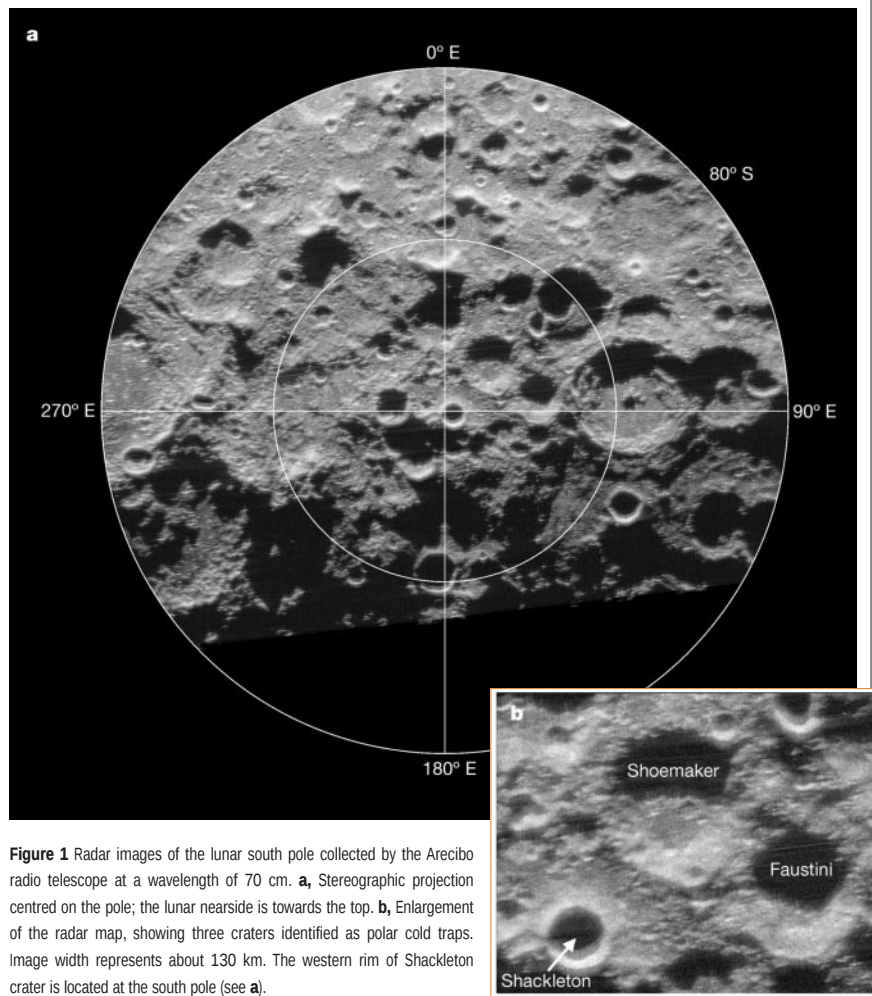


Figure 1 Radar images of the lunar south pole collected by the Arecibo radio telescope at a wavelength of 70 cm. **a**, Stereographic projection centred on the pole; the lunar nearside is towards the top. **b**, Enlargement of the radar map, showing three craters identified as polar cold traps. Image width represents about 130 km. The western rim of Shackleton crater is located at the south pole (see **a**).

permitted a viewing angle from the horizon of about 4° for the north pole and about 6° for the south pole (Fig. 1a). The limb of the Sun does not rise more than 2.1° above the horizon at the lunar poles, so the radar images show detail in regions that are permanently shadowed. The coded radar signal used to obtain these maps causes areas with high return, such as radar-facing crater walls, to blur along the delay axis (roughly parallel to the zero-longitude line in Fig. 1a). We therefore focused our study on crater floors, well away from the radar-bright walls.

Near the lunar north pole, the radar illuminates areas of permanent shadow in a crater close to 85° N, 63° E, the floor of Hermite crater, and several small craters within the large polar crater Peary. None of these areas seems to be different in radar-scattering behaviour from typical lunar highland terrain. Near the south pole, the floors of craters Shoemaker and Faustini, which are in permanent shadow¹, are partially illuminated

by the radar (Fig. 1b). Neither region gives rise to unusually strong radar echoes.

Lunar Prospector data⁵ suggest a high abundance of hydrogen for a shadowed¹, rugged intercrater area near 85° S, 310° E, but again the radar echoes are similar to those of other highland features. The radar-facing interior wall of Shackleton, some parts of which are in shadow, has a bright radar signature. The floor of Shackleton is not visible to the radar. The bright signature could arise from thick ice layers, but similarly high returns are observed from crater walls that do not lie in permanent solar shadow.

The absence of strong 70-cm backscatter from permanently shadowed crater floors, coupled with similar results at wavelengths of 3.5 cm and 12.6 cm, places strong constraints on the nature of putative water ice within the lunar regolith. Thick deposits of ice that are capable of coherent backscatter, which have been invoked to explain Clementine bistatic radar observations of the wall

of Shackleton⁴, are not observed within the crater floors visible to the Arecibo system. Any ice in these regions must be in the form of disseminated grains or thin (centimetres or less) interbedded layers, which could satisfy the Lunar Prospector results without strong radar backscatter enhancement.

This type of ice deposit, if present, would be considerably different from the thick, coherent layers observed in shadowed craters on Mercury^{7,8}. Such 'sparse' filling of the lunar cold-traps relative to Mercury could arise as a result of a lower average delivery rate of comets to the Moon, fortuitous recent comet impacts on Mercury, or a more rapid loss of ice on the lunar surface.

Bruce A. Campbell*, **Donald B. Campbell†**,
John F. Chandler‡, **Alice A. Hine§**,
Michael C. Nolan§, **Phillip J. Perillat§**

*Center for Earth and Planetary Studies,
Smithsonian Institution, Box 37012,
Washington DC 20013-7012, USA
e-mail: campbellb@nasm.si.edu

†Cornell University, Space Sciences Building,
Ithaca, New York 14853, USA

‡Smithsonian Astrophysical Observatory,
Cambridge, Massachusetts 02138, USA

§Arecibo Observatory, HC3 Box 53995,
Arecibo 00612, Puerto Rico

1. Margot, J. L., Campbell, D. B., Jurgens, R. F. & Slade, M. A. *Science* **284**, 1658–1660 (1999).
2. Bussey, D. B. J., Spudis, P. D. & Robinson, M. S. *Geophys. Res. Lett.* **26**, 1187–1190 (1999).
3. Nozette, S. et al. *Science* **274**, 1495–1497 (1996).
4. Nozette, S. et al. *J. Geophys. Res.* **106**, 23253–23266 (2001).
5. Feldman, W. C. et al. *J. Geophys. Res.* **106**, 23232–23252 (2001).
6. Black, G. J., Campbell, D. B. & Nicholson, P. D. *Icarus* **151**, 167–180 (2001).
7. Harmon, J. K. et al. *Nature* **369**, 213–215 (1994).
8. Black, G. J., Campbell, D. B. & Harmon, J. K. *Lunar Planet Sci. Conf. XXXIII* abstr. 1946 (2002).
9. Stacy, N. J. S., Campbell, D. B. & Ford, P. G. *Science* **276**, 1527–1530 (1997).

Competing financial interests: declared none.

Astrophysics

Refreshed shocks from a γ -ray burst

In addition to its remarkable supernova signature^{1,2}, the γ -ray burst of 29 March 2003 (GRB030329) had two interesting peculiarities: an unusually low-energy output in γ -rays and a large bump in its afterglow light curve after 1–2 days (followed by several less significant rebrightening episodes). We suggest that refreshed shocks — slow shells ejected from the source that catch up with the afterglow shock a relatively long time after the initial burst — produced the observed fluctuations in the early afterglow light curve and explain the low-energy output at early times.

The γ -ray emission in GRBs is thought to arise from internal shocks within a relativistic outflow from a compact source, which occur as different 'shells' within the outflow collide with each other (Fig. 1). At a greater distance, R , from the source, the ejecta decelerates as it

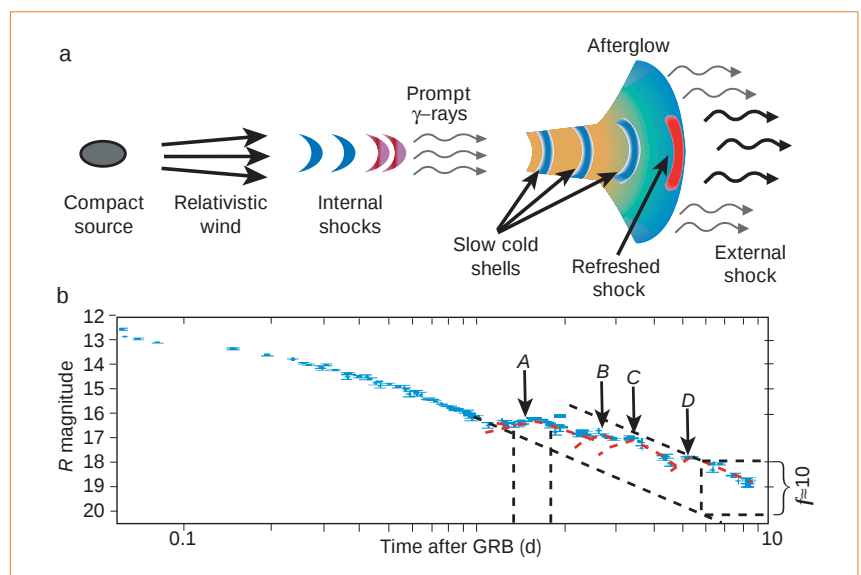


Figure 1 Refreshed shocks in γ -ray bursts. **a**, Illustration of a γ -ray burst: a compact source ejects a variable relativistic wind. Internal shocks within the outflow produce the γ -rays. At greater distances, the ejecta drives a strong shock into the surrounding medium, producing the afterglow. Slow shells ejected from the source catch up with the afterglow shock at late times, producing refreshed shocks; they thereby energize the afterglow and cause bumps in its light curve. **b**, Light curve of the GRB030329 burst (see <http://lanl.arxiv.org/abs/astro-ph/0304563> for refs) shows a large bump (A, red line) at $t \approx 1.3$ – 1.7 d, followed by three less significant bumps (B–D, dashed red lines, at $t \approx 2.4$ – 2.8 d, $t \approx 3.1$ – 3.5 d and $t \approx 4.9$ – 5.7 d, respectively).

drives a strong shock into the surrounding medium, producing the subsequent afterglow.

The achromatic increase in steepness of the optical light curve of GRB030329, from $t^{-\alpha}$, where t is the time measured from the burst, with $\alpha = \alpha_1 = 0.873 \pm 0.025$ at $t < t_j \approx 40.481 \pm 0.333$ days to $\alpha = \alpha_2 = 1.97 \pm 0.12$ at $t > t_j$ (ref. 3), resembled jet breaks seen in other GRBs. The inferred opening angle implies a prompt γ -ray energy output, E_γ , and X-ray luminosity at 10 h, L_X , that are factors of roughly 20 and 30, respectively, below the average values around which most GRBs are narrowly clustered³.

A well-monitored rebrightening with an amplitude $f \approx 2$ is evident at $t \approx 1.3$ – 1.7 days (Fig. 1), with a duration $\Delta t \approx t_j < t$. After the bump, the slope returns to $\alpha \approx \alpha_2$. Three additional, less significant bumps with similar features follow.

Angular smoothing generally suggests that light-curve variations should have $\Delta t \geq t$. However, here $\Delta t < t$. External density variations and patchy shells, which have been proposed to explain other variable afterglows, do not work here. The former requires an unrealistic external density; the latter fails because in GRB030329 the bumps occur after t_j , when the entire jet is visible.

Refreshed shocks (Fig. 1) arise when slow shells catch up with the afterglow shock at late times^{4–6}. Each collision causes a rebrightening in the afterglow light curve. After the rebrightening, the afterglow resumes its original decay slope. The stepwise shape of the light curve for GRB030329, where the same slope, α_2 , is regained after each bump, is a clear signature of refreshed shocks.

The collisions also inject energy into the afterglow shock. An energy increase of factor f increases the flux normalization by a factor $f = f^{(3+p)/4}$ for $v_m < v < v_c$ and $f = f^{(2+p)/4}$ for $v > v_m, v_c$, where p is the electron power-law index and $v_c(v_m)$ is the cooling (typical) frequency. For GRB030329, $p \approx 2$ and $v_c(t_j)$ is around the optical³, so f is roughly linear in f . The observed $f \approx 10$ (Fig. 1) implies $f \approx 10$, which brings the total energy close to the average value for all GRBs.

The original refreshed-shocks scenario⁵ predicts $\Delta t \approx t$. However, this assumes that refreshed shocks occur before t_j , whereas in GRB030329 they took place after t_j . As the later shells move in the wake of the forward shock, they remain cold and do not expand sideways. If all the shells have the same initial half-opening angle, θ_j , the duration of the rebrightening events will be $\Delta t \approx R\theta_j^2/2c = t_j(R/R_j) = t_j(t/t_j)^b$, where $b = 1/4$ ($b \approx 0$) if the lateral spreading of the jet is negligible (at the local sound speed). This allows $\Delta t < t \approx R/2\gamma^2 c$. Also, the width of the rear shell must be smaller than $c\Delta t$.

Numerical simulations suggest rather modest lateral spreading⁷, implying $b \approx 1/4$. This is consistent with $\Delta t \approx t_j$ for the first bump (and with $\Delta t \approx t_j, 2t_j$ for the later bumps). The energy increase after each refreshed shock causes a more rapid consequent increase in R ; the overall effect, however, is a factor ≤ 2 . The timing of the first bump suggests a Lorentz factor of about 6 for the slower shell.

Refreshed shocks can explain both the variability and the anomalously low values inferred for E_γ and L_X . A direct prediction of this interpretation is that there will be significant radio flares that correspond to the

observed optical bumps. These should arise from emission by the reverse shocks that form in the refreshed shocks.

Jonathan Granot*, Ehud Nakar†, Tsvi Piran†

*Institute for Advanced Study, Princeton, New Jersey 08540, USA

†Racah Institute for Physics, The Hebrew University, Jerusalem 91904, Israel

e-mail: udini@phys.huji.ac.il

1. Stanek, K. Z. *et al. Astrophys. J.* **591**, L17–L20 (2003).
2. Hjorth, J. *et al. Nature* **423**, 847–850 (2003).
3. Price, P. A. *et al. Nature* **423**, 844–847 (2003).
4. Rees, M. J. & Mészáros, P. *Astrophys. J.* **496**, L1–L4 (1998).
5. Kumar, P. & Piran, T. *Astrophys. J.* **532**, 286–293 (2000).
6. Sari, R. & Mészáros, P. *Astrophys. J.* **535**, L33–L37 (2000).
7. Granot, J., Miller, M., Piran, T., Suen, W. M. & Hughes, P. A. in *Gamma-Ray Bursts in the Afterglow Era* (eds Costa, E., Frontera, F. & Hjorth, J.) 312–314 (Springer, Berlin, 2001).

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Astrophysics

A constraint on canonical quantum gravity?

Gamma rays from the γ -ray burst (GRB) 021206 have been reported to be strongly linearly polarized¹, with the estimated degree of polarization ($80 \pm 20\%$) being close to the absolute maximum of 100% — affording us the opportunity to constrain models of quantum gravity, which has had 10^{10} years to act on the photons as they travelled towards us. Here I show that if the effects of quantum gravity are linearly proportional to the ratio of the photon energy to the characteristic scale energy of quantum gravity, then the polarization of photons with energies of about 0.1 MeV should be completely random, contrary to what is observed. I conclude that, should the polarization measurement be confirmed, quantum gravity effects act with a power that is greater than linearity, or that loop quantum gravity is not viable. Compared with previous methods and results (see ref. 2, for example), testing of the linear polarization of cosmic γ -ray bursts may substantially extend the observational window on the theory of quantum gravity.

GRBs are characterized by a highly variable flux of high-energy photons that propagate over cosmological distances. It has been suggested³ that they are the best candles in cosmological space, allowing us either to study or to constrain the effects of quantum gravity. These effects are known^{3,4} to be proportional to the ratio $(E/E_{\text{QG}})^n$ of the photon energy, E , to the Planck energy, $E_{\text{QG}} \approx 10^{19}$ GeV, and to the distance, D , of the photon's propagation. The linear case ($n = 1$) is the best studied^{3,4}, but the quadratic case ($n = 2$) has also recently been considered (see preprint at http://xxx.lanl.gov/PS_cache/gr-qc/pdf/0305/0305057.pdf). For $n = 1$, the effect of the energy-dependent refraction of

photons in quantum space-time should lead to a measurable difference in arrival time (of the order of milliseconds) for photons with different energies².

The linear polarization of γ -rays from GRB 021206 allows us to test another possible effect of quantum space-time, which is predicted for canonical quantum gravity in loop representation. In this case, space-time exhibits the property of birefringence⁴: two photons with opposite states of helicity, $+1$ and -1 , have different group velocities

$$v_{\pm} = c(1 \pm \chi(E/E_{\text{QG}})^n) \quad (1)$$

The factor χ is about 1 for loop representation of quantum gravity³. A linearly polarized electromagnetic wave may be represented as the superposition of two monochromatic waves with opposite circular polarizations. When a linearly polarized wave propagates inside the substance with birefringence, the plane of polarization rotates along the path because of the difference in group velocity between the two circular components.

For the linear case $n = 1$, the phase angle, φ , of a plane of linear polarization changes along a distance D (in light years) as

$$\Delta\varphi_1(E) \approx \chi(D/hc)E^2/E_{\text{QG}}^2 \approx 10^4 \chi(E/0.1 \text{ MeV})^2 D \quad (2)$$

This angle depends on the photon energy as E^2 . Linear polarization measured within a broad energy range should vanish, provided that the difference in accumulated angles is large for photons with different energies. Two photons with energies of around 0.1 MeV and with a difference of energy of about 0.01% will therefore accumulate a difference of $\delta\varphi \approx \chi$ in polarization phase angle after a year of propagation in space with birefringence (see equation (2)).

For cosmological GRBs, which have a travel distance of $D \approx 10^{10}$ light years, the planes of linear polarization of photons with different energies should be totally randomized. The bulk linear polarization of photons with energies greater than 0.01 eV over a broad energy range must become zero even if they were all originally 100% polarized in a single plane.

In the quadratic case of quantum space-time birefringence with $n = 2$, the rotation of a plane of linear polarization is rather small for photons with energy of around 0.1 MeV

$$\Delta\varphi_2(E) \approx \chi(D/hc)E^3/E_{\text{QG}}^2 \approx 10^{-19} \chi(E/0.1 \text{ MeV})^3 D \quad (3)$$

However, the distance $D \approx 10^{10}$ light years is so large that even the quadratic case of birefringence could be tested by polarization measurements of photons with energies greater than 100 MeV. The detection⁵ of a high-energy component of GRB941017 (energies up to 200 MeV), which dominates the total fluence of the event, suggests that

quadratic space-time birefringence could be tested experimentally in the future by polarimetry of such GRBs.

I therefore conclude that either the birefringence of quantum space-time with $n = 1$ should be below the level of $\chi > 10^{-14}$, or it should be quadratic ($n = 2$), assuming that the strong linear polarization of GRBs is confirmed by a second measurement.

Igor G. Mitrofanov

Space Research Institute, Profsojuznaya str. 84/32, 117997 Moscow, Russia

e-mail: imitrofa@space.ru

1. Colburn, W. & Boggs, S. E. *Nature* **423**, 415–417 (2003).
2. Jacobson, T., Liberati, D. & Mattingly, D. *Nature* **424**, 1019–1021 (2003).
3. Amelino-Gamella, G., Ellis, J., Mavromatos, N. E., Nanopoulos, D. V. & Sarkar, S. *Nature* **393**, 763–765 (1998).
4. Gambini, R. & Pullin, J. *Phys. Rev. D* **59**, 124021 (1999).
5. Gonzalez, M. M. *et al. Nature* **424**, 749–751 (2003).

COMMUNICATIONS ARISING

Condensed-matter physics

Spurious magnetism in high- T_c superconductor

One challenge in condensed-matter physics is to unravel the interplay between magnetism and superconductivity in copper oxides with a high critical temperature (T_c). Kang *et al.*¹ claim to have revealed a quantum phase transition from the superconducting to an antiferromagnetic state in the electron-doped material $\text{Nd}_{2-x}\text{Ce}_x\text{CuO}_4$ (NCCO) based on the observation of magnetic-field-induced neutron-scattering intensity at $(1/2, 1/2, 0)$, $(1/2, 0, 0)$ and related reflections. Here we argue that the observed magnetic intensity is due to a secondary phase of $(\text{Nd,Ce})_2\text{O}_3$. We therefore contend that the effect is spurious and not intrinsic to superconducting NCCO.

To achieve superconductivity in NCCO, a rather severe oxygen-reduction procedure has to be applied². We have discovered that the reduction process decomposes a small (0.01–0.10%) volume fraction of NCCO. The resultant $(\text{Nd,Ce})_2\text{O}_3$ secondary phase has the complex cubic bixbyite structure, common among rare-earth (RE) sesquioxides³, with a lattice constant, a_c , that is about $2\sqrt{2}$ times the planar lattice constant of tetragonal NCCO. $(\text{Nd,Ce})_2\text{O}_3$ is epitaxial with the host lattice, with long-range order parallel to the CuO_2 planes of NCCO, but extending only about $5a_c$ perpendicular to the planes. Because of the relationship between the two lattice constants, certain structural reflections from the impurity phase appear at seemingly commensurate NCCO positions — that is, the cubic $(2, 0, 0)$, reflection can also be indexed as $(1/2, 1/2, 0)$. However, there is roughly a 10% mismatch between a_c and the c -lattice constant of NCCO, and therefore $(0, 0, 2)_c$ can also be indexed as $(0, 0, 2.2)$.

There are 32 rare-earth ions in the RE_2O_3 unit cell, belonging to two crystallographically distinct sites with inequivalent saturated moments³. At the $(2,0,0)_c$ reflection, the contributions from the two rare-earth sites interfere destructively, which should lead to a peak in the observed scattering intensity in the paramagnetic phase if the moments saturate at different fields. Although the magnetic structure and spin hamiltonian of epitaxial, quasi-two-dimensional $(\text{Nd,Ce})_2\text{O}_3$ are unknown, it is possible to devise simple experiments to test whether the field-induced scattering is due to NCCO or $(\text{Nd,Ce})_2\text{O}_3$.

Kang *et al.* find that at a temperature of 5 K, the $(1/2,1/2,0)$ (that is, $(2,0,0)_c$) intensity reaches a peak at a field of about 6.5 T, and argue that this peak is associated with the upper critical field B_{c2} of NCCO. Figure 1a summarizes the field dependence of an $x=0.18$ superconducting sample of ours in the temperature range 1.9–10 K. Our data agree with those of Kang *et al.* The figure shows that the intensity scales with B/T and exhibits a peak consistent with two-moment paramagnetism. Furthermore, as the upper critical field of a superconductor increases with decreasing temperature, this implies that the reported correspondence of the peak position with B_{c2} at 5 K is coincidental. We do not observe spontaneous neodymium ordering of either $(\text{Nd,Ce})_2\text{O}_3$ or NCCO down to 1.4 K.

Figure 1b, c shows that the field effects reported by Kang *et al.* are also observable in a non-superconducting, oxygen-reduced, $x=0.10$ sample, both at the previously reported positions and at positions that are unrelated to the NCCO lattice but equivalent in the cubic lattice of $(\text{Nd,Ce})_2\text{O}_3$. Not only are the incommensurate positions $(0,0,2,2)$ and $(1/4,1/4,1,1)$ unrelated to the proposed NCCO magnetic order, but the physical situation of the magnetic field applied parallel (in the cases of the $(0,0,2,2)$ and $(1/4,1/4,1,1)$) or perpendicular (in all other cases) to the CuO_2 planes is fundamentally different in that the upper critical fields for the two geometries differ significantly. Note that $(1/2,0,0)$ and $(1/4,1/4,1,1)$ correspond to $(1,1,0)_c$ and $(1,0,1)_c$, respectively. Care was taken to ensure that in all cases the magnetic field was applied along a cubic axis of $(\text{Nd,Ce})_2\text{O}_3$ and perpendicular to the scattering wavevector.

These simple experimental tests demonstrate that the observed field effects in oxygen-reduced NCCO result from an epitaxial secondary phase of $(\text{Nd,Ce})_2\text{O}_3$.

P. K. Mang*, S. Laroche†, M. Greven*‡

*Department of Applied Physics, †Department of Physics, and ‡Stanford Synchrotron Radiation Laboratory, Stanford University, Stanford, California 94305, USA

e-mail: greven@stanford.edu

1. Kang, H. J. *et al.* *Nature* **423**, 522–525 (2003).

2. Tokura, Y., Takagi, H. & Uchida, S. *Nature* **337**, 345–347 (1989).

3. Moon, R. M., Koehler, W. C., Child, H. R. & Raubenheimer, L. J. *Phys. Rev.* **176**, 722–731 (1968).

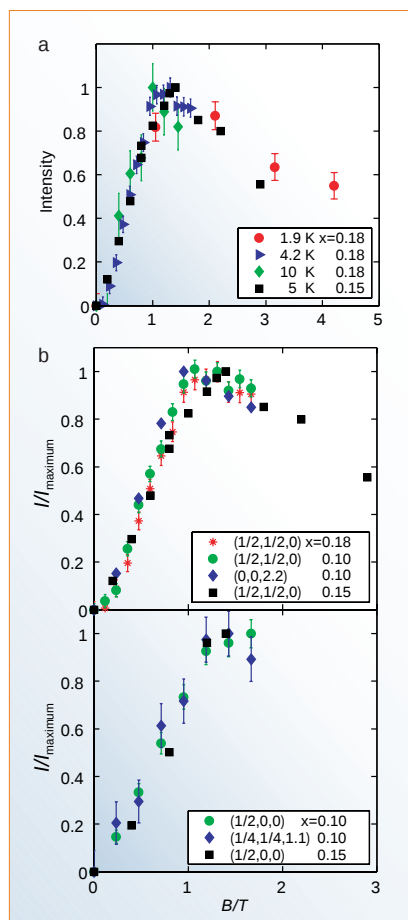


Figure 1 Field and temperature dependence of magnetic scattering. **a**, Arbitrarily scaled scattering intensity at $(1/2,1/2,0)$ for a superconducting sample of NCCO (nominal cerium concentration $x=0.18$; $T_c=20$ K) as a function of B/T with the field along $[0,0,1]$. The results are compared with the data of Kang *et al.*¹ ($x=0.15$; $T=5$ K). **b**, **c**, Comparison of the results of Kang *et al.* with data taken at $T=4$ K for a superconducting sample ($x=0.18$) and a non-superconducting sample ($x=0.10$). Superconductivity in NCCO can be achieved only for $x>0.13$. The magnetic field is applied along $[1,1,0]$ for $(0,0,2,2)$ and $(1/4,1/4,1,1)$ and along $[0,0,1]$ in all other cases. Data were normalized by maximum intensity. Full details are available from the authors.

Kang et al. reply — Mang *et al.* observe a cubic $(\text{Nd,Ce})_2\text{O}_3$ impurity phase grown epitaxially in annealed samples of electron-doped $\text{Nd}_{2-x}\text{Ce}_x\text{CuO}_4$ (NCCO). They claim that this impurity phase has long-range order parallel to the CuO_2 planes of NCCO but extending only about $4a_c$ perpendicular to the planes, thus forming a quasi-two-dimensional $(\text{Nd,Ce})_2\text{O}_3$ lattice matched with the a – b plane of NCCO.

Although we have confirmed the presence of such an impurity phase, $(\text{Nd,Ce})_2\text{O}_3$ in our samples forms a three-dimensional long-range structural order¹ and is unrelated to the quasi-two-dimensional superlattice reflections^{1,2}. In the paramagnetic state of $(\text{Nd,Ce})_2\text{O}_3$, a field will induce a net moment on magnetic Nd. By arbitrarily scaling the impurity scattering at $(0,0,2,2)$ for fields less

than 7 T to our c -axis field-induced scattering of NCCO at $(1/2,1/2,0)$, Mang *et al.* argue that our observed magnetic scattering² is due entirely to $(\text{Nd,Ce})_2\text{O}_3$. We disagree, however.

There are three ways to resolve this impurity problem. First, if the magnetic scattering at $(1/2,1/2,0)$ (ref. 2) is due entirely to $(\text{Nd,Ce})_2\text{O}_3$, one would expect the field-induced intensity to be identical when B is parallel to the c -axis and when it is parallel to the $[1,-1,0]$ axis, as required by the cubic symmetry of $(\text{Nd,Ce})_2\text{O}_3$. Experimentally, we find that the field-induced effect at $(1/2,1/2,0)$ is much larger when B is parallel to the c -axis¹, which is inconsistent with the cubic symmetry of $(\text{Nd,Ce})_2\text{O}_3$ but consistent with the upper critical field of NCCO being much smaller in this geometry^{1,2}.

Second, as the lattice parameter of $(\text{Nd,Ce})_2\text{O}_3$ does not match the c -axis lattice parameter of NCCO (ref. 1), measurements at non-zero integer L cannot be contaminated by $(\text{Nd,Ce})_2\text{O}_3$. Our experiments indicate that the $(1/2,1/2,3)$ peak shows an induced antiferromagnetic component when the field is along the c -axis and hence superconductivity is strongly suppressed¹, but not when in the a – b plane and superconductivity is only weakly affected². This is direct proof of the connection between field-induced antiferromagnetic order and suppression of superconductivity in NCCO. We also note that the qualitatively different behaviour observed when B is perpendicular to c , in comparison with when it is parallel to c , directly violates the cubic symmetry of $(\text{Nd,Ce})_2\text{O}_3$.

Finally, an independent report³ confirms our principal findings^{1,2} in studies of annealed superconducting $\text{Pr}_{0.89}\text{LaCe}_{0.11}\text{CuO}_4$ (PLCCO), a similar electron-doped material in which the cubic impurity phase $(\text{Pr,L a,C e})_2\text{O}_3$ has a non-magnetic ground state and no field dependence below 7 T (our unpublished observations). For fields up to 5 T, Fujita *et al.*³ find enhanced antiferromagnetic order at $(1/2,3/2,0)$ with increasing field in PLCCO. Above 5 T, this order decreases with increasing field, which is consistent with the field dependence of $(1/2,1/2,0)$ of NCCO (ref. 2). The agreement between two different electron-doped systems in two independent experiments^{1–3} confirms the quantum phase transition from the superconducting to an antiferromagnetic state in electron-doped, high- T_c superconductors².

H. J. Kang, Pengcheng Dai*, J. W. Lynn, M. Matsuura, J. R. Thompson, Shou-Cheng Zhang, D. N. Argyriou, Y. Onose, Y. Tokura

*Department of Physics and Astronomy, University of Tennessee, Knoxville, Tennessee 37996-1200, and Condensed Matter Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6393, USA

e-mail: daip@ornl.gov

1. Matsuura, M. *et al.* *Phys. Rev. B* **68**, 144503 (2003).

2. Kang, H. J. *et al.* *Nature* **423**, 522–525 (2003).

3. Fujita, M., Matsuuda, M., Katano, S. & Yamada, K. *Physica B* (in the press).

Hyperelasticity governs dynamic fracture at a critical length scale

Markus J. Buehler^{1*}, Farid F. Abraham^{2*} & Huajian Gao^{1*}

¹Max Planck Institute for Metals Research, Heisenbergstrasse 3, 70569 Stuttgart, Germany

²IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120, USA

* These authors contributed equally to this work

The elasticity of a solid can vary depending on its state of deformation. For example, metals will soften and polymers may stiffen as they are deformed to levels approaching failure. It is only when the deformation is infinitesimally small that elastic moduli can be considered constant, and hence the elasticity linear. Yet, many existing theories model fracture using linear elasticity, despite the fact that materials will experience extreme deformations at crack tips. Here we show by large-scale atomistic simulations that the elastic behaviour observed at large strains—hyperelasticity—can play a governing role in the dynamics of fracture, and that linear theory is incapable of fully capturing all fracture phenomena. We introduce the concept of a characteristic length scale for the energy flux near the crack tip, and demonstrate that the local hyperelastic wave speed governs the crack speed when the hyperelastic zone approaches this energy length scale.

Why and how cracks spread in brittle materials is of essential interest to numerous scientific disciplines and technological applications, and a theoretical understanding is essential for many engineering applications. We show by large-scale atomistic simulation that hyperelasticity, the elasticity of large strains, can play a governing role in the dynamics of brittle fracture. This is in contrast to many existing theories of dynamic fracture where the linear elastic behaviour of solids is assumed to be sufficient to predict materials failure^{1–3}. Real solids have elastic properties that are significantly different for small and for large deformations.

Many phenomena associated with rapidly propagating cracks are not thoroughly understood. Some experimental work^{4,5} as well as many computer simulations^{6–8} have shown a significantly reduced crack propagation speed in comparison with the predictions by the theory. In contrast, other experiments indicated that over 90% of the Rayleigh wave speed can be achieved^{9–15}. Such discrepancies between theories, experiment and simulations cannot always be attributed to the fact that real solids have all sorts of imperfections such as grain boundaries and microcracks (either pre-existing or created during the crack propagation), because similar discrepancies also appear in molecular-dynamics simulations of cracks travelling in perfect atomic lattices.

Gao^{16,17} and Abraham^{7,18,19} have independently proposed that hyperelastic effects at the crack tip may play an important role in the dynamics of fracture. Their suggestions have been used to help interpret phenomena related to crack branching and dynamic crack tip instability, as well as explaining the significantly lower maximum crack propagation speed observed in some experiments and many computer simulations. However, it is not generally accepted that hyperelasticity should play a significant role in dynamic fracture. One reason for this belief stems from the fact that the zone of large deformation in a loaded body with a crack is highly confined to the crack tip, so that the region where linear elastic theory does not hold is extremely small compared to the extensions of the specimen^{1,2}.

In this study, we use large-scale molecular-dynamics simulations^{19–24} in conjunction with continuum mechanics concepts^{1,2} to prove that hyperelasticity can be crucial for understanding dynamic fracture. Our study shows that local hyperelasticity around the crack tip can significantly influence the limiting speed of cracks by enhancing or reducing local energy flow. This is true even if the zone of hyperelasticity is small compared to the specimen dimensions. The hyperelastic theory completely changes the concept of the

maximum crack velocity in the classical theories. For example, the classical theories clearly predict that mode I cracks are limited by Rayleigh wave speed and mode II cracks are limited by longitudinal wave speed. In contrast, both super-Rayleigh mode I and supersonic mode II cracks are allowed by hyperelasticity and have been seen in computer simulations^{23,25}. In our simulations, we find that there exists a characteristic length scale associated with energy flow near the crack tip such that hyperelasticity completely dominates crack dynamics if the size of the hyperelastic region approaches this characteristic length. In earlier simulations^{23,25}, a nonlinear inter-atomic stiffening was assumed, and there was no sharp distinction between the linear and nonlinear elastic regimes for the stretched solid. In contrast, our model is based on a biharmonic potential composed of two spring constants, one associated with small deformations and the other associated with large deformations. This serves as a simplistic model material for hyperelasticity, allowing us to investigate the generic features of hyperelasticity common to a large class of real materials.

Modelling

We consider propagation of a crack in a two-dimensional simulation geometry shown in Fig. 1. The slab size is given by l_x and l_y .

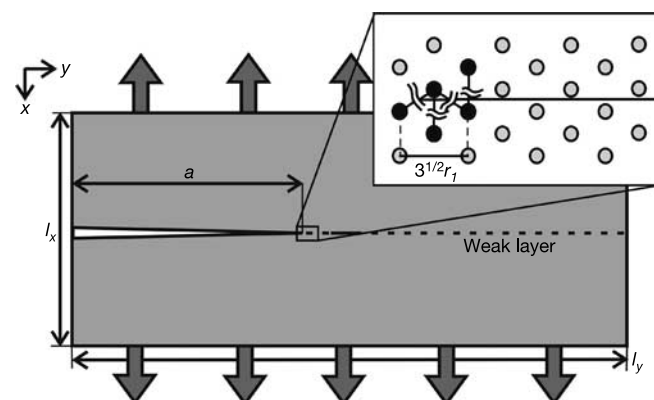


Figure 1 The simulation geometry with the lattice orientation and the weak layer. The crack propagates in the y direction, and the slab is loaded in mode I in the x direction. A slit of length $a = 200$ serves as an initial crack.

The crack propagates in the y direction, and its extension is denoted by a . The crack propagates in a triangular hexagonal lattice with nearest-neighbour distance $r_1 = 2^{1/6} \approx 1.12246$ along the crystal orientation shown in Fig. 1. To avoid crack branching, a weak fracture layer is introduced by assuming that atomic bonds across the prospective crack path snap at a critical atomic snapping distance r_{break} while those in the rest of the slab never break. The snapping distance can be used to adjust the fracture surface energy γ .

We adopt a biharmonic, interatomic potential composed of two spring constants $k_1 = 36/\sqrt[3]{2} \approx 28.57$ and $k_2 = 2k_1$ (all quantities given are in dimensionless units). We consider two ‘model materials’, one with elastic stiffening and the other with elastic softening behaviour. In the elastic stiffening system, the spring constant k_1 is associated with small perturbations from the equilibrium distance r_1 , and the second spring constant k_2 is associated with large bond stretching for $r > r_{\text{on}}$. The role of k_1 and k_2 is reversed in the elastic softening system ($k_1 = 2k_2$, and $k_2 = 36/\sqrt[3]{2}$). Purely harmonic systems are obtained if r_{on} is chosen to be larger than r_{break} .

Poisson’s ratio is found to be approximately independent of strain and around $\nu \approx 0.33$ for all potentials. In the stiffening system, the small deformation (up to about 0.5% of strain) Young’s modulus is $E_1 = 33$ with shear modulus $\mu = 12.4$, and the large deformations tangent Young’s modulus is $E_2 = 66$ with shear modulus $\mu_2 = 24.8$. The values are reversed for the softening system where the small deformation Young’s modulus is $E_1 = 66$, and the large deformation tangent Young’s modulus is $E_2 = 33$. The two-dimensional sheet of atoms behaves as an isotropic plane stress sheet at small deformation.

There are three distinct wave velocities in elastic solids: the longitudinal wave speed $c_l = \sqrt{3\mu/\rho}$, the shear wave speed $c_s = \sqrt{\mu/\rho}$ and the Rayleigh wave speed $c_r \approx 0.9225c_s$, with the density $\rho = 2/\sqrt{3}/\sqrt[3]{2} \approx 0.9165$ for atomic mass $m = 1$. It follows from the

choice of $k_2 = 2k_1$ that there is a ratio of $\sqrt{2}$ between small and large deformation wave speed.

Crack speed and energy flow

We show by molecular-dynamics simulations that a localized, small hyperelastic region around the crack tip can have significant effects on the dynamics of crack propagation.

In all simulations, the slab is statically loaded with 0.32% strain in mode I. The strain energy density far ahead of the crack tip is given by $S = 0.5\varepsilon_{xx}^2 J_x E / (1 - \nu^2)$, where E is the Young’s modulus at small strain. The linear elastic expression of strain energy density is valid because material far ahead of the crack is strained always below the onset threshold of the bilinear law, that is, it remains in the linear elastic regime of material response. The strain and strain energy density both vanish far behind the crack. For a unit distance of crack propagation, a strip of material with energy density S ahead of the crack is replaced by an identical strip with zero strain energy behind the crack. According to the linear elastodynamic theory of fracture¹, the crack speed should satisfy the dynamic energy release rate equation $A(v/c_r) = 2\gamma/S$ where the function $A(v/c_r)$ is a universal function of crack velocity for a given material. Assuming that the small-strain elasticity completely governs the dynamics of fracture, the linear theory predicts that crack velocity should depend only on the ratio S/γ . During crack propagation, the energy stored ahead of the crack tip is partly converted by the bond-breaking process into fracture surface energy, and partly dissipated into atomic motion.

In the purely harmonic case, the fracture surface energy γ depends on r_{break} and E . In the biharmonic case, the fracture surface energy depends on r_{break} , r_{on} , E_1 and E_2 . Our strategy is to focus on the prediction from linear theory that crack velocity depends only on the ratio S/γ . To achieve this objective, we keep the ratio S/γ constant in all of the simulations. In the harmonic systems, as $S \propto E$ and $\gamma \propto E$ (see Methods section), we choose the parameter r_{break} to be identical in all cases. In the biharmonic systems, we adjust the parameter r_{break} at given values of r_{on} , E_1 and E_2 , to always keep S/γ constant.

We choose $r_{\text{break}} = 1.17$ for the harmonic systems. The failure strain at the crack tip can reach a magnitude of several per cent, which is comparable to many ‘real materials’. In the harmonic systems (with Young’s modulus equal to E_1 or E_2), the crack achieves the same propagation velocity of around 80% of the Rayleigh wave speed. This is consistent with the linear theory.

For the biharmonic systems, we choose $r_{\text{on}} = 1.1275$ and $r_{\text{break}} = 1.1558$ in the stiffening system and $r_{\text{break}} = 1.1919$ in the softening system to keep S/γ constant. In contrast to the linear theory prediction, we find that the crack propagation velocity is about 20% larger in the stiffening system and 30% smaller in the softening system. These deviations cannot be explained by the linear theory. The fact that we change the large-strain elasticity while keeping the small-strain elasticity constant indicates that hyperelasticity is affecting crack dynamics.

A geometric criterion based on the principal strain is used to characterize the area with hyperelastic material response close to the crack tip. The region occupied by atoms having a local maximum principal strain $\epsilon_1 > (r_{\text{on}} - r_1)/r_1$ defines the hyperelastic area A . Figure 2a shows the hyperelastic area in the case of a stiffening material, and Fig. 2b shows the hyperelastic area in the case of an elastically softening material, indicating that the hyperelastic effect is highly localized to the crack tip (these pictures show a portion of the simulation slab near the crack tip). However, the effect of hyperelasticity on crack velocity is significant, independent of the slab size.

A measure for the direction and magnitude of energy flow in the vicinity of the crack tip is the dynamic Poynting vector^{1,26}. The magnitude of the dynamic Poynting vector, $P = \sqrt{P_1^2 + P_2^2}$, may be identified as a measure for the local energy flow. A measure for the enhancement or reduction of energy flow is obtained by subtracting

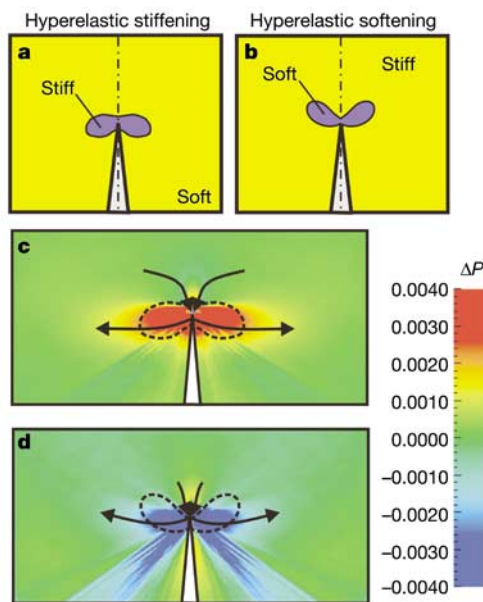


Figure 2 Hyperelastic zone and the associated change in energy flow near the crack tip. **a, b**, The hyperelastic zones for the bilinear stiffening system and the softening system, respectively. **c, d**, The associated change of energy flow ΔP in the stiffening and softening systems. The local hyperelastic zone is indicated in purple in **a** and **b** and by a dotted line in **c** and **d**. Plots **a** and **b** show a section of $1,150 \times 930$ (16% of the total simulation slab area), plots **c** and **d** show a section of $1,150 \times 575$ (10% of the total simulation slab area).

Table 1 Change of energy flow towards the crack tip

	Change of net energy flow to crack tip	Change of energy flow towards crack tip	Change of energy flow away from crack tip	Change of limiting speed
Stiffening	+19%	+20%	+25%	+20%
Softening	−32%	−32%	−35%	−30%

The table summarizes change of net energy flow, change of energy flow towards the crack tip and change of energy flow away from the crack tip, in comparison to the harmonic system.

the magnitude of the dynamic Poynting vector in the harmonic case from that in the biharmonic case at every point in the slab, $\Delta P = P = P_{\text{biharm}} - P_{\text{harm}}$. If the difference is negative, energy flow is reduced, and if the difference is positive, energy flow is enhanced. The steady-state fields are averaged over space as well as time to obtain good statistics.

Figure 2 shows the energy flow enhancement and reduction in the vicinity of the crack tip for the elastically stiffening bilinear system (Fig. 2c) and for the elastically softening system (Fig. 2d). In each plot, the local hyperelastic zone is indicated by a dotted line. The energy flow in the vicinity of the crack tip is enhanced in the bilinear stiffening case and reduced in the softening case. In these plots, we also indicate the direction of energy flow with arrows and note that in the softening case, the energy flow ahead of the crack almost vanishes. The plots show that the local hyperelastic effect leads to an enhancement (stiffening system) or reduction (softening system) in energy flow. The small hyperelastic regions enhance the energy flow around the crack tip. The higher crack velocity in the stiffening system and the lower velocity in the softening system are due to enhancement or reduction of the energy flow in the vicinity of the crack tip.

Table 1 summarizes change of net energy flow, as well as change of energy flow towards and away from the crack tip, in comparison to the harmonic system. The results quantify those depicted in Fig. 2c and d and show that the net energy flow as well as the flow of energy towards and away from the crack tip are all enhanced in the stiffening case, and reduced in the softening case. The integral of energy flux, or path-independent dynamic J -integral, around the crack tip increases by 19% in comparison with the harmonic case for the stiffening system, while decreasing by 32% for the softening system.

How fast can cracks propagate?

We have learned that a local hyperelastic zone around the crack tip can have significant effect on the velocity of the crack. For a mode I tensile crack, linear theory predicts that the energy release rate vanishes for all velocities in excess of the Rayleigh wave speed^{1–3}, implying that a mode I crack cannot move faster than the Rayleigh wave speed. This prediction is indeed confirmed in systems with the harmonic potential where crack velocity approaches the Rayleigh wave speed independently of the slab size, provided that the applied strain is larger than 1.08% and the slab width is sufficiently large ($l_x > 1,000$). The systems are loaded dynamically in this case. Our strain levels are about ten times lower than in many other studies²⁴.

We consider hyperelastic effect of different strengths by using a biharmonic potential with different onset strains governed by the parameter r_{on} . The parameter r_{on} governs the onset strain of the hyperelastic effect $\epsilon_{\text{on}} = (r_{\text{on}} - r_1)/r_1$. The simulations reveal crack propagation at super-Rayleigh velocities in steady-state with a local stiffening zone around the crack tip. Figure 3a plots the crack velocity as a function of the hyperelasticity onset strain ϵ_{on} . The crack speeds shown in Fig. 3 are determined during steady-state propagation. We observe that the earlier the hyperelastic effect is turned on, the larger the limiting velocity. Measuring the hyperelastic area using the principal strain criterion, we find that the area grows as ϵ_{on} becomes smaller. A correlation of the square root of the hyperelastic area with the achieved limiting speed of the crack is shown in Fig. 3a. In Fig. 3b, we depict the shape of the hyperelastic area near the crack tip for different choices of ϵ_{on} . The shape and size

of the hyperelastic region is found to be independent of the slab width l_x . In all cases, the hyperelastic area remains confined to the crack tip and does not extend to the boundary of the simulation. Figure 3 shows that the hyperelastic effect is very sensitive to the potential parameter and the extension of the local hyperelastic zone.

Mode I cracks can travel at steady-state intersonic velocities if there exists a locally stiffening hyperelastic zone. For example, when the large-strain spring constant is chosen to be $k_2 = 4k_1$, with $r_{\text{on}} = 1.1375$ and $r_{\text{break}} = 1.1483$ (that is, ‘stronger’ stiffening and thus larger local wave velocity than before), the mode I crack propagates 21% faster than the Rayleigh speed of the soft material, and becomes intersonic, as shown by the Mach cone of shear wave front depicted in Fig. 4a.

We have also simulated a shear-dominated mode II crack using the biharmonic stiffening potential. We define $r_{\text{break}} = 1.17$, and r_{on} is chosen slightly below r_{break} to keep the hyperelastic region small. The dynamic loading is stopped soon after the daughter crack is nucleated^{25,27,28}. The result is shown in Fig. 4b. The daughter crack nucleated from the mother crack propagates supersonically through the material, although the hyperelastic zone remains localized to the crack tip region. Supersonic mode II crack propagation has been observed previously^{23,25} using an anharmonic stiffening potential. However, a clearly defined hyperelastic zone could not be specified in those simulations. Our result proves that a local hyperelastic stiffening effect at the crack tip causes supersonic crack propagation, in contrast to the linear continuum theory.

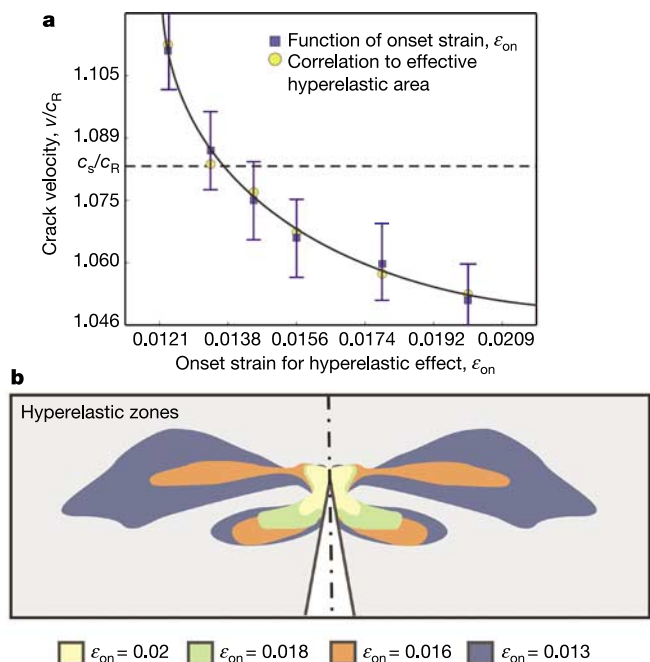


Figure 3 Change of crack speed with the onset strain of hyperelasticity. The crack speed as a function of the onset strain of hyperelasticity ϵ_{on} , as well as a correlation of the crack speed with the square root of the hyperelastic area is shown in **a**. In **b**, the shape of the hyperelastic zone near the crack tip is shown for different onset strains, indicating that the hyperelastic area remains confined to the vicinity of the crack tip. **b**, A section of $1,650 \times 570$ (9% of the total simulation slab area).

The observation of super-Rayleigh and intersonic mode I cracks, as well as supersonic mode II cracks, clearly contradicts the prediction by the classical theories of fracture.

Characteristic energy length scale

The problem of a super-Rayleigh mode I crack in an elastically stiffening material is somewhat analogous to Broberg's²⁹ problem of a mode I crack propagating in a stiff elastic strip embedded in a soft matrix. The geometry of this problem is shown in Fig. 5a. Broberg²⁹ has shown that, when such a crack propagates supersonically with respect to the wave speeds of the surrounding matrix, the energy release rate can be expressed in the form

$$G = \frac{\sigma^2 h}{E} f(v, c_1, c_2) \quad (1)$$

where σ is the applied stress, h is the half-width of the stiff layer and f is a non-dimensional function of crack velocity and wave speeds in the strip and the surrounding matrix (c_1, c_2). The dynamic Griffith energy balance requires $G = 2\gamma$, indicating that crack propagation velocity is a function of the ratio h/χ where $\chi \propto \gamma E/\sigma^2$ can be defined as a characteristic length scale for local energy flux. By dimensional analysis, the energy release rate of our hyperelastic stiffening material is expected to have similar features except that Broberg's strip width h should be replaced by a characteristic size of the hyperelastic region r_H . Therefore, we introduce the concept of a characteristic length

$$\chi = \beta \frac{\gamma E}{\sigma^2} \quad (2)$$

for local energy flux near a crack tip. The coefficient β may depend on the ratio between hyperelastic and linear elastic properties. We have simulated the Broberg problem and found that the mode I crack speed reaches the local Rayleigh wave speed as soon as h/χ reaches unity. Numerous simulations verify that the scaling law in equation (2) holds when γ, E and σ is changed independently. The results are shown in Fig. 5b. From the simulations, we estimate numerically $\beta \approx 87$ and $\chi \approx 750$.

The existence of a characteristic length for local energy flux near the crack tip has not been discussed in the literature and plays the central role in understanding the effect of hyperelasticity. Under a

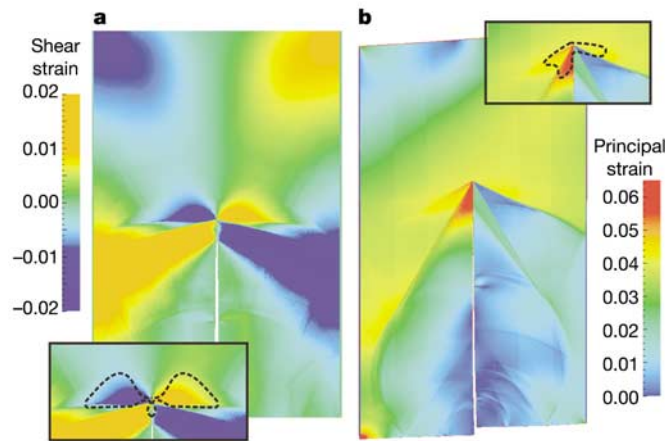


Figure 4 Intersonic mode I crack motion and supersonic mode II crack motion. **a**, The shear strain field in a section of the slab during intersonic mode I crack propagation. The Mach cone is associated with the shear wave speed. The hyperelastic zones are indicated in the insets by dotted lines. **b**, The principal strain field of a crack propagating at a supersonic velocity in mode II. The two Mach cones associated with shear wave and longitudinal wave can clearly be recognized. These plots illustrate that cracks can move with super-Rayleigh speeds under mode I loading and supersonic speeds under mode II loading. The classical theory of fracture cannot explain these results.

particular experimental or simulation condition, the relative importance of hyperelasticity is determined by the ratio r_H/χ . For small r_H/χ , the crack dynamics is dominated by the global linear elastic properties because much of the energy transport necessary to sustain crack motion occurs in the linear elastic region. However, when r_H/χ approaches unity, as is the case in some of our molecular-dynamics simulations, the dynamics of the crack is dominated by local elastic properties because the energy transport required for crack motion occurs within the hyperelastic region.

The concept of energy characteristic length χ immediately explains how the classical barrier for transport of energy over large distances can be undone by rapid transport near the tip.

Discussion and conclusions

We have shown that local hyperelasticity has a significant effect on the dynamics of brittle crack speeds and have discovered a characteristic length associated with energy transport near a crack tip. The assumption of linear elasticity fails if there is a hyperelastic zone in the vicinity of the crack tip comparable to the energy characteristic length. Therefore, we conclude that hyperelasticity is crucial for understanding and predicting the dynamics of brittle fracture. Our simulations prove that even if the hyperelastic zone extends only a small area around the crack tip, there may be crucial effects on the limiting speed and the energy flow towards the crack tip, as illustrated in Fig. 3. If there is a local softening effect, we find that the limiting crack speed is lower than in the case of harmonic solid.

Our study has shown that hyperelasticity dominates the energy transport process when the hyperelastic zone size becomes comparable to the characteristic length $\chi \propto \gamma E/\sigma^2$. Under normal experimental conditions, the magnitude of stress may be one or

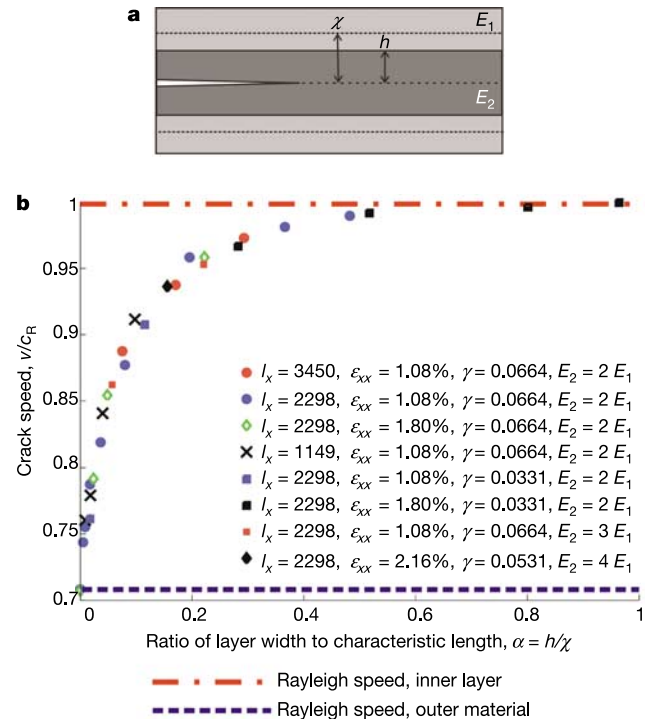


Figure 5 Scaling law for the characteristic length associated with energy transport near the crack tip. **a**, Broberg's problem. **b**, The simulation results of crack velocity versus the half-width h of Broberg's stiff layer normalized by the characteristic $\chi = \beta \gamma E/\sigma^2$. The simulation results from a variety of different simulation cases (for example, variation of surface energy, Young's modulus, loading) confirm the scaling law for the characteristic length of energy transport near the crack tip. Once h reaches the characteristic length, the crack velocity reaches the Rayleigh velocity of the inner layer.

two orders of magnitude smaller than that under molecular-dynamics simulations. In such cases, the characteristic length χ is relatively large and the effect of hyperelasticity on effective velocity of energy transport is relatively small. However, χ decreases with the square of the applied stress. At about 1% of elastic strain as in our simulations, this zone is already on the order of a few hundred atomic spacing and significant hyperelastic effects are observed.

Our simulations indicate that the universal function $A(v/c_R)$ in the classical theory of dynamic fracture is no longer valid once the hyperelastic zone size r_H becomes comparable to the energy characteristic length χ . Linear elastic fracture mechanics predicts that the energy release rate of a mode I crack vanishes for all velocities in excess of the Rayleigh wave speed. However, this is only true if $r_H/\chi \ll 1$. A hyperelastic theory of dynamic fracture should incorporate this ratio into the universal function so that the function should be generalized as $A(v/c_R, r_H/\chi)$. The local hyperelastic zone changes not only the near-tip stress field within the hyperelastic region, but also induces a finite change in the integral of energy flux around the crack tip. We find that the dynamic J -integral around a super-Rayleigh (Fig. 3) mode I crack is still path-independent but no longer vanishes in the presence of hyperelasticity. Similarly, the supersonic mode II crack motion (Fig. 4b) can only be understood from the point of view of hyperelasticity. A single set of global wave speeds is not capable of capturing all phenomena observed in dynamic fracture.

We believe that the length scale χ , heretofore missing in the existing theories of dynamic fracture, will prove to be helpful in forming a comprehensive picture of crack dynamics. In most engineering and geological applications, typical values of stress are much smaller than those in molecular-dynamics simulations. In such cases, the ratio r_H/χ is small and the effective speed of energy transport is close to predictions by linear elastic theory. However, the effect of hyperelasticity will be important for nanoscale materials, such as highly strained thin films or nanostructured materials, as well as high-speed impact phenomena. □

Methods

Molecular dynamics simulation procedure

Our simulation tool is classical molecular dynamics^{23,30}. Molecular dynamics predicts the motion of a large number of atoms governed by their mutual interatomic interaction, and it requires numerical integration of the equations of motion, usually via a Verlet algorithm with time step $\Delta t = 0.0036$ (refs 23, 30) (in reduced time unit). The slab size is chosen large enough such that waves reflected from the boundary do not interfere with the propagating crack. We establish a linear velocity gradient before simulation to avoid shock wave generation from the boundaries. To strain the system, we use two approaches. The first is using a constant strain rate applied over a loading time t , by displacing the outermost rows of atoms. After the loading time, the boundaries are kept fixed. In the second method, we strain the system before simulation in the loading direction, and keep the boundary fixed during simulation. In either way, the crack starts to move once a critical strain is applied. It can be shown that the stress intensity factor remains constant in a strip geometry inside a region of $3/4l_x < a < (l_y - 3/4l_x)$ (refs 24, 31). This ensures that the crack achieves a steady-state during propagation through the slab. The slab is initialized at zero temperature before simulation.

Slab geometry and typical length scales

The length l_y is several times larger than l_x with the ratio l_y/l_x ranging from two to five. The slab width l_x considered ranges from 1,150 (smallest) up to 4,597 (largest), corresponding to micrometre length scale in physical dimensions). The largest model contains over 70 million atoms. All quantities in this paper are given in reduced units. The condition for small-scale yielding is satisfied in all cases (with harmonic, stiffening and softening potentials), because breaking of atomic bonds occurs over a region involving only a few atoms along the weak layer (that is, a very small fracture process zone). There is no dislocation process and the system is perfectly brittle. The slab is loaded with a maximum of a few per cent strain, according to the crack loading mode. The loading is significantly lower than other studies²⁴. A slit of length $a = 200$ is cut midway through the slab as an initial, atomically sharp crack.

Interatomic potential

The interatomic potential is defined as

$$\phi(r) = \begin{cases} 1/2k_1(r - r_1)^2 & \text{if } r < r_{on} \\ a_2 + 1/2k_2(r - r_2)^2 & \text{if } r \geq r_{on} \end{cases} \quad (3)$$

The other parameters of the potential are determined to be $a_2 = 1/2k_1(r_{on} - r_1)^2 - 1/2k_2(r_{on} - r_2)^2$, and $r_2 = 1/2(r_{on} + r_1)$ from continuation conditions. For a harmonic two-dimensional elastic sheet of atoms, Young's modulus is given by $E = \frac{2}{\sqrt{3}}k_1$, and the shear modulus is $\mu = \frac{\sqrt{3}}{4}k_1$ (see for example, ref. 16).

Crack tip velocity measurements

Accurate determination of crack tip velocity is important because we need to be able to measure even the smallest changes in the propagation speed. The crack tip position is determined by finding the surface atom with maximum y position in the interior of a search region inside the slab. This quantity is averaged over a small time interval to eliminate very-high-frequency fluctuations. To obtain the steady-state velocity of the crack, the measurements are taken within a region of constant stress intensity factor^{24,31}. In addition to checking the velocity history, steady-state is verified by path-independency of the energy flux integral¹.

Fracture surface energy

The fracture surface energy γ is defined as the energy necessary to break atomic bonds per unit crack advance, and is given by

$$\gamma = \frac{E(r_{break} - r_1)^2}{2r_1} \quad (4)$$

for the purely harmonic potential and

$$\gamma = \frac{2a_2 + k_2(r_2 - r_{break})^2}{\sqrt{3}r_1} = \frac{E_1(r_{on} - r_1)^2 - E_2((r_{on} - r_2)^2 - (r_2 - r_{break})^2)}{2r_1} \quad (5)$$

for the biharmonic potential (if $r_{break} > r_{on}$), where E_1 is the Young's modulus associated with small deformations, and E_2 is the Young's modulus associated with large deformations.

Atomic strain

We define the left Cauchy–Green strain tensor of an atom l (ref. 32)

$$b_{ij} = \left\{ \frac{1}{3r_l^2} \sum_{k=1}^N [(x_i^l - x_j^l)(x_j^l - x_k^l) + (x_j^l - x_k^l)(x_k^l - x_i^l) + (x_k^l - x_i^l)(x_i^l - x_j^l)] \right\} \quad (6)$$

with $N = 6$ nearest neighbours. Unlike the virial stress, the virial strain is valid instantaneously in space and time. The maximum principal strain b_1 is obtained by diagonalization of the strain tensor b_{ij} . The maximum principal engineering strain is then given by $\epsilon_1 = \sqrt{b_1} - 1$.

Energy flow analysis

Under steady-state conditions, the net energy flow to the crack tip is defined by an integration of the product of dynamic Poynting vector with surface normal n_j over a contour $\partial\Omega$ surrounding the crack tip, (the contour could be selected as a circle around the crack tip)¹

$$F_N(\partial\Omega) = \int_{\partial\Omega} \{ \sigma_{ij} n_j \dot{u}_i + (U + T) v n_i \} \cdot d\Omega \delta_i \quad (7)$$

Here, the dynamic Poynting vector $P_j = \sigma_{ij} \dot{u}_i + (U + T) v \delta_{ij}$ measures the local energy flow, with stress σ_{ij} , particle velocity $\dot{u}_i$, potential energy density U , kinetic energy density T , and crack propagation speed v .

Computing techniques and supercomputing centre

A FORTRAN molecular-dynamics code parallelized with MPI is used for the simulations. We use the IBM XL FORTRAN compiler for AIX for compilation of the source code. The simulations are carried out on IBM Power 4 Regatta nodes in the Max Planck Society Supercomputer Centre in Munich.

Received 17 April; accepted 19 September 2003; doi:10.1038/nature02096.

- Freund, L. B. *Dynamic Fracture Mechanics* 2nd edn (Cambridge Univ. Press, Cambridge, UK, 1998).
- Broberg, B. *Cracks and Fracture* (Academic, San Diego, 1999).
- Slepyan, L. I. *Models and Phenomena in Fracture Mechanics* (Springer, Berlin, 2002).
- Fineberg, J., Gross, S. P., Marder, M. & Swinney, H. L. Instability in dynamic fracture. *Phys. Rev. Lett.* **67**, 141–144 (1991).
- Ravi-Chandar, K. Dynamic fracture of nominally brittle materials. *Int. J. Fract.* **90**, 83–102 (1998).
- Abraham, F. F., Brodbeck, D., Rudge, W. E. & Xu, X. Instability of fracture—a computer-simulation investigation. *Phys. Rev. Lett.* **73**, 272–275 (1994).
- Abraham, F. F., Brodbeck, D. & Rudge, W. E. A molecular-dynamics investigation of rapid fracture mechanics. *J. Mech. Phys. Solids* **45**, 1595–1619 (1997).
- Falk, M. L., Needleman, A. & Rice, J. R. A critical evaluation of cohesive zone models of dynamic fracture. *J. Phys. IV France* **11**, 43–50 (2001).
- Washabaugh, P. D. & Knauss, W. G. A reconciliation of dynamic crack velocity and Rayleigh-wave speed in isotropic brittle solids. *Int. J. Fract.* **65**, 97–114 (1994).
- Sharon, E. & Fineberg, J. Confirming the continuum theory of dynamic brittle fracture for fast cracks. *Nature* **397**, 333 (1999).
- Field, J. E. Brittle fracture—its study and application. *Contemp. Phys.* **12**, 1–31 (1971).
- Cramer, T., Wanner, A. & Gumbsch, P. Energy dissipation and path instabilities in dynamic fracture of silicon single crystals. *Phys. Rev. Lett.* **85**, 788–791 (2000).
- Hauch, J. A. & Marder, M. P. Energy balance in dynamic fracture, investigated by a potential drop technique. *Int. J. Fract.* **90** (1–2), 133–151 (1998).
- Hauch, J. A., Holland, D., Marder, M. P. & Swinney, H. L. Dynamic fracture in single crystal silicon. *Phys. Rev. Lett.* **82**, 3823–3826 (1999).
- Hull, D. & Beardmore, P. Velocity of propagation of cleavage cracks in Tungsten. *Int. J. Fract. Mech.* **2**, 468–488 (1966).

16. Gao, H. A theory of local limiting speed in dynamic fracture. *J. Mech. Phys. Solids* **44**, 1453–1474 (1996).
17. Gao, H. Elastic waves in a hyperelastic solid near its plane-strain equibiaxial cohesive limit. *Phil. Mag. Lett.* **76**, 307–314 (1997).
18. Abraham, F. F. Dynamics of brittle fracture with variable elasticity. *Phys. Rev. Lett.* **77**, 272–275 (1996).
19. Abraham, F. F. *et al.* Instability dynamics in three-dimensional fracture: An atomistic simulation. *J. Mech. Phys. Solids* **45**, 1461–1471 (1997).
20. Rountree, C. L. *et al.* Atomistic aspects of crack propagation in brittle materials: Multimillion atom molecular-dynamics simulations. *Annu. Rev. Mater. Res.* **32**, 377–400 (2002).
21. Zhou, S. J., Lomdahl, P. S., Thomson, R. & Holian, B. L. Dynamic crack processes via molecular-dynamics. *Phys. Rev. Lett.* **76**, 2318–2321 (1996).
22. Heitzler, S. I., Kessler, D. A. & Levine, H. Mode I fracture in a nonlinear lattice with viscoelastic forces. *Phys. Rev. E* **6**, 016126 (2002).
23. Abraham, F. F. *et al.* Simulating materials failure by using up to one billion atoms and the world's fastest computer: Brittle fracture. *Proc. Natl Acad. Sci. USA* **99**, 5777–5782 (2002).
24. Swadener, J. G., Baskes, M. I. & Nastasi, M. Molecular dynamics simulation of brittle fracture in silicon. *Phys. Rev. Lett.* **89**, 085503 (2002).
25. Abraham, F. F. The atomic dynamics of fracture. *J. Mech. Phys. Solids* **49**, 2095–2111 (2001).
26. Fratini, S., Pla, O., Gonzalez, P., Guinea, F. & Louis, E. Energy radiation of moving cracks. *Phys. Rev. B* **66**, 104104 (2002).
27. Abraham, F. F. & Gao, H. How fast can cracks propagate? *Phys. Rev. Lett.* **84**, 3113–3116 (2000).
28. Rosakis, A. J. Intersonic shear cracks and fault ruptures. *Adv. Phys.* **51**, 1189–1257 (2002).
29. Broberg, K. B. Dynamic crack propagation in a layer. *Int. J. Sol. Struct.* **32** (6–7), 883–896 (1995).
30. Allen, M. & Tildesley, D. *Computer Simulation of Liquids* (Oxford Univ. Press, 1989).
31. Knauss, W. G. Stresses in an infinite strip containing a semi-infinite crack. *J. Appl. Mech.* **33**, 356–362 (1966).
32. Zimmermann, J. *Continuum and Atomistic Modeling of Dislocation Nucleation at Crystal Surface Ledges* PhD thesis, Stanford Univ. (1999).

Acknowledgements F.F.A. acknowledges the generous support provided by the Humboldt Research Award for Senior US Scientists.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.G. (hjgao@mf.mpg.de).

Efflux-dependent auxin gradients establish the apical–basal axis of *Arabidopsis*

Jiří Friml^{1*}, Anne Vieten^{1*}, Michael Sauer^{1*}, Dolf Weijers^{1,2}, Heinz Schwarz³, Thorsten Hamann^{1†}, Remko Offringa² & Gerd Jürgens¹

¹Zentrum für Molekularbiologie der Pflanzen, Universität Tübingen, Auf der Morgenstelle 3, 72076 Tübingen, Germany

²Institute of Biology, Clusius Laboratory, Leiden University, Wassenaarseweg 64, 2333 AL, Leiden, The Netherlands

³Max Planck Institute for Developmental Biology, Spemannstrasse 35, D-72076 Tübingen, Germany

* These authors contributed equally to this work

† Present address: Department of Plant Biology, Carnegie Institution of Washington, 260 Panama street, 94305 Stanford, USA

Axis formation occurs in plants, as in animals, during early embryogenesis. However, the underlying mechanism is not known. Here we show that the first manifestation of the apical–basal axis in plants, the asymmetric division of the zygote, produces a basal cell that transports and an apical cell that responds to the signalling molecule auxin. This apical–basal auxin activity gradient triggers the specification of apical embryo structures and is actively maintained by a novel component of auxin efflux, PIN7, which is located apically in the basal cell. Later, the developmentally regulated reversal of PIN7 and onset of PIN1 polar localization reorganize the auxin gradient for specification of the basal root pole. An analysis of *pin* quadruple mutants identifies PIN-dependent transport as an essential part of the mechanism for embryo axis formation. Our results indicate how the establishment of cell polarity, polar auxin efflux and local auxin response result in apical–basal axis formation of the embryo, and thus determine the axiality of the adult plant.

The development of higher eukaryotes includes the generation of a species-specific body plan. As a first step, a multicellular organization is established from a single-celled zygote during embryogenesis, with cells adopting specific fates according to their relative positions. In *Drosophila*, maternal organizers initiate a cascade of spatially restricted transcription factors that partitions the anterior–posterior axis of the embryo¹. In *Caenorhabditis elegans*, sperm entry triggers anterior–posterior axis specification, initiating a series of asymmetric cell divisions that establish founder cells with different fate potential². In plants, the mature embryo displays a main axis of polarity, with the shoot meristem flanked by the cotyledons (embryonic leaves) at the top end and separated by hypocotyl (embryonic stem) and root from the root meristem at the opposite pole. The origin of this apical–basal pattern, which is remarkably uniform across flowering plant species, has been traced back to early embryogenesis in *Arabidopsis*³. The zygotic division generates a smaller apical and a larger basal cell. The apical cell divides vertically and eventually gives rise to all apical embryo structures. The basal cell continues to divide horizontally and produces the suspensor, which connects the embryo with the maternal tissue³. The uppermost suspensor cell is subsequently recruited by the embryo and specified to become the hypophysis—the founder of the root meristem. At the triangular stage, the apical pole of the embryo is organized with the initiation of two symmetrically positioned cotyledons.

Several indirect lines of evidence implicated the plant hormone auxin (indole-3-acetic acid, IAA) in embryo development. Embryo patterning mutants, such as *monopteros* (*mp*), *bodenlos* (*bdl*) or *gnom* (*gn*)^{4,5}, lack basal structures and display variably fused cotyledons. Molecular analysis of these mutants has revealed that *MP* and *BDL* encode the transcriptional activator auxin response factor 5 (ARF5) and the corresponding transcriptional repressor IAA12, respectively, both of which are involved in auxin response^{6,7}, and that *GN* encodes a regulator of vesicle trafficking that mediates the subcellular targeting of auxin-transport components^{8,9}. For later development, chemical manipulation of auxin distribution has

suggested a link between patterning and auxin transport^{10–13}. Moreover, in wheat and carrot later-stage embryos, auxin has been detected using a microscale technique^{14,15}.

Auxin is actively distributed within the plant by efflux-dependent cell-to-cell movement¹⁶. The direction of auxin flow was proposed to be determined by the asymmetric cellular localization of efflux carriers^{17,18}, probably represented by plant-specific PIN proteins¹⁹. However, reported *pin* mutants^{13,20} as well as other auxin transport or response mutants display only mild and infrequent early embryonic defects. Moreover, the presence of auxin or its response has not been demonstrated in early embryogenesis. Thus, the function of auxin in apical–basal pattern formation during embryogenesis of higher plants is debatable, and the process of early pattern formation itself is mechanistically only poorly understood.

Here we show dynamic gradients of auxin accumulation and response during *Arabidopsis* early embryogenesis, which are mediated by cellular efflux and required for apical–basal axis formation. A novel auxin efflux regulator, PIN7, and other functionally redundant PIN proteins are important determinants of both auxin gradients and apical–basal axis establishment. Our results suggest a model of how spatially separated auxin transport and response from the zygotic division onwards, together with the establishment of cell polarity, mediate patterning along the initial apical–basal auxin gradient.

Apical–basal auxin gradients in embryogenesis

The activity of the synthetic auxin-responsive promoter *DR5* has been used to visualize the spatial pattern of auxin response, and hence indirectly the distribution of auxin^{12,13,21,22}. We constructed a fluorescent variant, *DR5rev::GFP*, which enabled us to monitor auxin response and its dynamics from the earliest stages of embryogenesis onwards. Immediately after the division of the zygote, *DR5* activity was pronounced in the smaller, apical cell (Fig. 1a). Signal intensity rapidly increased in the developing proembryo, whereas only very weak signal was detected in the suspensor (Fig. 1b, c). This apical–basal auxin response gradient was reversed at around the

32-cell stage: the maximum of *DR5* activity shifted basally into the uppermost suspensor cells, including the hypophysis (Fig. 1d, e). At later stages of embryogenesis, additional *DR5* signals appeared in the tips of the developing cotyledons and in the provascular strands (Fig. 1f). When the quiescent centre of the root meristem was established, the *DR5* maximum shifted further basally into the adjacent columella precursors (Fig. 1f, inset). *DR5* activity persisted in these cells, showing the pattern previously reported for post-embryonic development^{12,13,23}.

Several controls were performed to examine whether *DR5rev::GFP* acted as a reliable reporter not only for auxin response but also for cellular auxin levels. First, exogenously supplied auxin induced *DR5* activity in all embryo cells (Fig. 2d), suggesting that the spatially restricted signals in untreated embryos reflected differences in auxin levels between cells. Second, another *DR5* variant and another reporter—the diphtheria toxin (DTA), a highly sensitive, non-fluorescent reporter that causes cell lethality²⁴—were used. DTA expression by the *DR5* promoter was made conditional through the *GAL4/UAS* transactivation system²⁵. In control experiments, β -glucuronidase (*GUS*) activity in *DR5>>GUS* embryos was undetectable before the late globular stage (not shown), whereas experiments with other activator lines revealed *UAS* transcription to be activated with variable strength and timing in sibling embryos^{26,27}. This variation in both onset and strength of *GAL4/UAS* transactivation enabled us to assess DTA effects at different

stages of embryogenesis. DTA expression before the globular stage caused mainly symptoms of ablation in apical cells, as visualized by the enhanced nuclear fluorescence of Schiff staining (Fig. 1g). By contrast, later DTA expression led exclusively to defects in the basal region (Fig. 1h, compare with Fig. 4a). Thus, the *DR5>>DTA* reporter displayed the same spatiotemporal activity pattern as the *DR5rev::GFP* reporter. Finally, we monitored the accumulation of auxin itself by immunolocalization with an anti-IAA antibody²⁸. Increased IAA levels were detected apically in the pre-globular embryo (Fig. 1i), whereas the basal part of the embryo stained strongly from the globular stage onwards (Fig. 1j). Later, the IAA maximum moved from the quiescent centre of the root meristem into the columella precursors (Fig. 1j, inset), as observed for *DR5* (see Fig. 1f, inset). Thus, the dynamic IAA accumulation pattern mirrored the *DR5* activity pattern, confirming the apical–basal reversal of auxin distribution and response during early embryogenesis.

Cellular efflux links auxin gradients and patterning

Experimental manipulation of auxin homeostasis in the tiny *Arabidopsis* embryo has not been possible so far. We established *in vitro* culture of embryos within excised ovules. Cultured embryos developed normally and displayed the stage-specific patterns of *DR5* activity (Fig. 2f). To examine whether auxin transport has a role in *DR5* activity distribution, embryos were cultured in the presence of synthetic auxins, the efflux substrate 1-naphthaleneacetic acid (NAA) or the influx substrate 2,4-dichlorophenoxyacetic acid (2,4-D)²⁹. NAA increased *DR5* activity without changing its pattern (Fig. 2a, c). By contrast, 2,4-D interfered with the *DR5* pattern, resulting in a strong signal in the entire embryo (Fig. 2d). Inhibiting auxin efflux either by the phytochrome NPA or the vesicle-trafficking inhibitor brefeldin A (BFA) perturbed the *DR5* pattern. At early

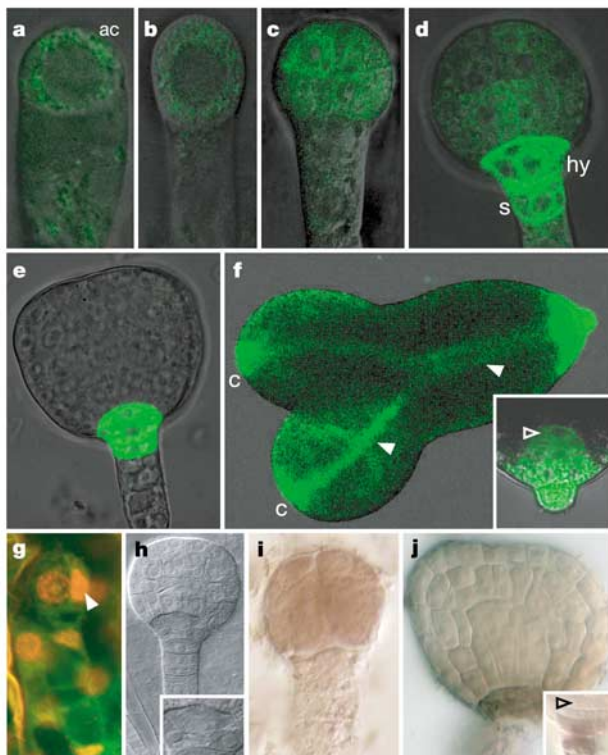


Figure 1 Auxin and auxin response in embryogenesis. **a–f**, *DR5rev::GFP* auxin response. Maximum in the apical cell (ac) lineage after zygotic division (**a**), and at the one-cell (**b**) and the eight-cell (**c**) stage. Shift to the hypophysis (hy) and upper suspensor (s) cells in young globular (**d**) and triangular (**e**) embryos. Additional signals in cotyledon (c) tips and provascular (arrowheads) at the torpedo stage (**f**); inset, maximum below the quiescent centre (open arrowhead). **g, h**, *DR5>>DTA* expression. Apical defects in young embryos (**g**); arrowhead, dying proembryo cell. Basal defects in globular embryo (**h**). **i, j**, IAA accumulation in a 16-cell-stage proembryo (**i**), at the basal pole of a triangular embryo (**j**), and below the quiescent centre (open arrowhead) at the torpedo stage (inset). GFP fluorescence in green (**a–f**). Schiff staining in red (**g**). IAA signals in brown (**i, j**).

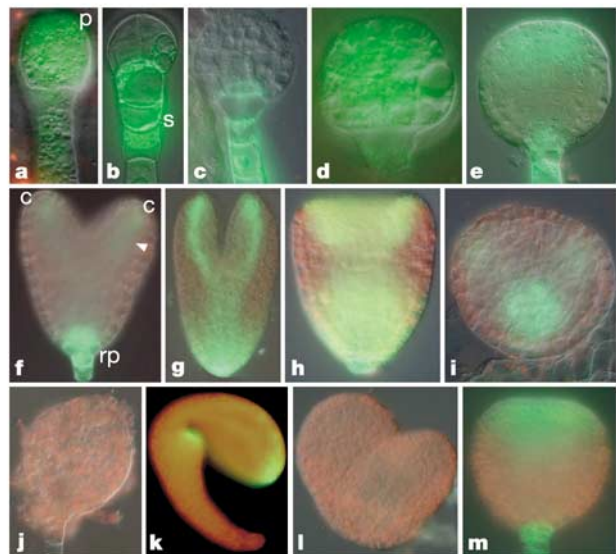


Figure 2 *DR5* auxin response in *in vitro* cultured ovules and mutants. **a, b**, Pre-globular embryos treated with: NAA, enhanced signal in proembryo (p) (**a**); NPA, ectopic suspensor (s) signal (**b**). **c–e**, Globular embryos treated with: NAA, enhanced basal signal (**c**); 2,4-D, ubiquitous signal (**d**); BFA, ectopic apical signal (**e**). **f**, Untreated heart-stage embryo. *DR5* signals at the root pole (rp), cotyledon (c) tips and weaker in provascular tissue (arrowhead). **g–i**, Long-term treatment with: NAA, enhanced signal but no change in spatial pattern (**g**); 2,4-D, abnormal *DR5* activity, cotyledon and root pole specification compromised (**h**); BFA, in extreme cases, no apical–basal axis establishment (**i**). **j–m**, No basal signal in *mp* (**j, k**) or *bdl* (**l**) embryos, occasional signal in developing cotyledon tips (**k**), ectopic apical signal in *gn* embryos (**m**). Times of culture: 16 h (**a–e**), 3 days (**f–i**). GFP fluorescence in green. Autofluorescence in red.

stages, GFP (green fluorescent protein) fluorescence was detected in the suspensor (Fig. 2b, compare with Fig. 2a) rather than in the proembryo, whereas after reversal of the gradient, an ectopic *DR5* maximum appeared in the embryo apex (Fig. 2e, compare with Fig. 1d, e). These data strongly suggest that the auxin distribution in embryos, as reflected by *DR5* activity, is mediated by auxin efflux.

Long-term interference with auxin homeostasis enabled us to assess the relationship between auxin gradients and embryo patterning. NAA had no adverse effects (Fig. 2g). By contrast, the auxin analogue 2,4-D (Fig. 2h) as well as auxin-efflux inhibitors NPA and BFA (Fig. 2i) caused abnormalities in auxin distribution, which were always accompanied by embryo defects, ranging from cup-shaped embryos with misspecified apical structures and a non-functional root pole, to ball-shaped embryos without any discernible apical–basal axis. These defects resembled the phenotypes of *gn* and shared some features, such as lack of a functional root, with *mp* and *bdl*. In *mp* and *bdl* mutants, neither the early apical *DR5* activity nor the later basal *DR5* maximum was detectable (Fig. 2j, l). Only occasionally did older embryos show GFP expression at the tips of cotyledons (Fig. 2k), suggesting an involvement of other components of auxin response late in embryogenesis. *gn* mutants displayed a severely altered *DR5* activity pattern at the globular stage. Both the ectopic *DR5* maximum in the apical part of the embryo and the abnormally positioned peak in the suspensor mirrored the *DR5* activity pattern upon auxin-efflux inhibition (Fig. 2m, compare to Fig. 2e). Thus, both drug treatments and the analysis of embryo mutants correlated apical–basal patterning and cotyledon specification with spatial patterns of auxin distribution and response.

PIN expression in embryogenesis

To identify the molecular components of auxin efflux that mediate the dynamic auxin gradients in embryogenesis, we analysed the expression and function of PIN regulators of auxin efflux. Four out of eight *Arabidopsis* PIN genes, *PIN1*–*PIN4*, have been described¹⁶. We isolated full-length complementary DNAs for the novel members *PIN5*–*PIN8*. By sequence analysis, *PIN5* and *PIN8* are divergent family members that lack important domains and may have other functions. For the other six, we analysed the expression during embryogenesis using *PIN::GUS* transgenic plants, *in situ* hybridization and/or immunolocalization experiments. On the basis of *promoter::GUS* studies, neither *PIN2* nor *PIN6* expression was detectable in embryos (not shown). *PIN1::GUS* activity as well as *PIN1* messenger RNA was detected in the apical cell lineage (Fig. 3a). The earliest activity of *PIN3::GUS* was detected at the basal pole of the heart-stage embryo, confirmed by *PIN3* mRNA localization in the precursors of the columella (Fig. 3f). *PIN4* protein localized to the descendants of the hypophysis and to provascular initials of the root meristem (Fig. 3g), as previously reported¹³. The novel gene *PIN7* displayed an expression pattern complementary to that of *PIN1*. A *Ds-GUS* enhancer-trap insertion³⁰ marked the basal cell lineage from the zygotic division onwards (Fig. 3h), which was confirmed for later stages by the expression pattern of a *PIN7::GUS* translational fusion (Fig. 3i) and by *PIN7* mRNA *in situ* hybridization (Fig. 3j). Thus, at least four PIN genes are expressed during embryogenesis, providing a system for regulated auxin distribution.

PIN polarity correlates with auxin distribution

Asymmetric subcellular localization of PIN proteins has been correlated with the direction of auxin flow in postembryonic development¹⁶. We therefore examined whether *PIN1* and *PIN7* localization correlated with the apical–basal auxin gradients in early embryogenesis. From the one-cell to the 16-cell stage, *PIN1* marked all newly formed cell boundaries within the proembryo, without any detectable polarity (Fig. 3b). At the 32-cell stage, *PIN1* became

polarly localized in the provascular cells facing the basal embryo pole—the hypophysis (Fig. 3c). This event coincided with the basal shift of the auxin response maximum to the hypophysis (see Fig. 1d). Within the forming root meristem, *PIN1* also shifted to the basal side of the quiescent centre cells (Fig. 3d, e), which again coincided with the shift of the auxin maximum to the basally adjacent columella precursors (see Fig. 1f, j, insets).

To immunolocalize *PIN7* protein in the embryo, we raised anti-*PIN7* antiserum. *PIN7* was detected in the basal cell immediately after the zygotic division, both in endomembranes and at the boundary facing the smaller apical cell (Fig. 3k), which was complementary to the auxin response maximum in the apical cell (see Fig. 1a). Until the 32-cell stage, *PIN7* continued to reside at the apical side of suspensor cells facing the developing proembryo (Fig. 3l, m). Thus, *PIN7* localization reflects early polarization of basal cells, representing the earliest polarity marker known. At around the 32-cell stage, the asymmetric localization of *PIN7* suddenly reversed, shifting to the basal side of suspensor cells (Fig. 3n) for the rest of embryogenesis (Fig. 3o). After formation of the lens-shaped cell, *PIN7* localized to all its boundaries (Fig. 3o, inset). Most interestingly, onset of the basal accumulation of *PIN1* in the proembryo cells and reversal of *PIN7* polarity correlated with the apical-to-basal reversal of the auxin gradient.

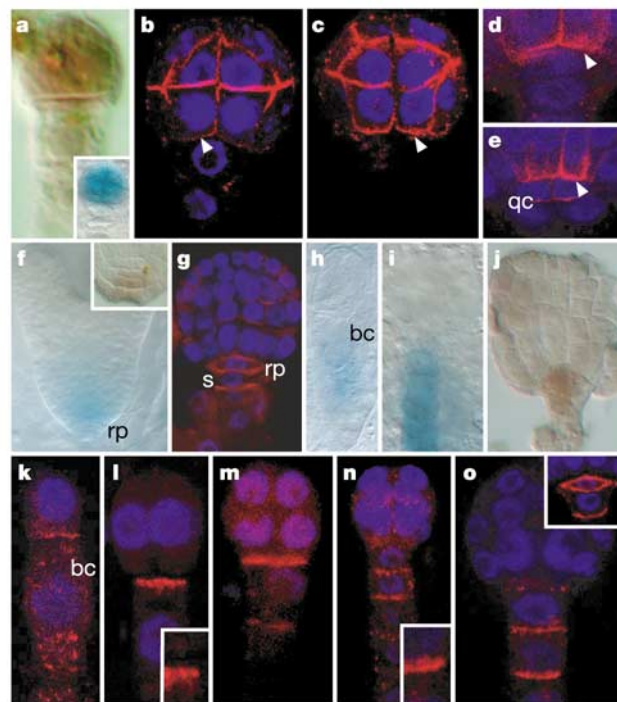


Figure 3 PIN expression and protein localization in embryogenesis. **a**, *PIN1* mRNA in a two-cell-stage proembryo, and *PIN1::GUS* (inset) in an eight-cell-stage proembryo. **b–e**, *PIN1* protein at the inner cell boundaries of a 16-cell-stage proembryo (**b**), basally in provascular initials (arrowheads) of 16/32-cell-stage (**c**) and globular (**d**) embryos, and additional signals in the quiescent centre (qc) at the heart stage (**e**). **f**, *PIN3::GUS* and *PIN3* mRNA (inset) at the root pole (rp) of a heart-stage embryo. **g**, *PIN4* protein at the basal end of provascular initials, at the root pole and in the uppermost suspensor (s) cell of a globular embryo. **h–j**, *PIN7* expression in the basal cell (bc) lineage. *Ds-GUS* enhancer-trap line, one-cell-stage (**h**); *PIN7::GUS*, globular embryo (**i**); *PIN7* mRNA, triangular embryo (**j**). **k–o**, *PIN7* protein in the basal cell (bc) of a one-cell embryo (**k**); apically in the basal cell of two-cell (**l**) and eight-cell (**m**) embryos; basally in suspensor cells of 16/32-cell (**n**) and globular (**o**) embryos, and around a lens-shaped cell (inset). mRNA signals in brown (**a**, inset **f**, **j**). Protein signals in red and DAPI nuclear counterstain in blue (**b–e**, **g**, **k–o**). *GUS* staining in blue (**a**, inset **f**, **h**, **i**).

Apical–basal pattern defects in *pin* mutants

To investigate the biological function of *PIN7*, we isolated three mutant alleles with insertions within the coding region (*pin7-1*, -2 and -3) and generated *PIN7* RNAi (RNA interference) transgenic plants. RNA and protein analyses confirmed that the mutant alleles were null (not shown). The temporal and spatial correlation of *PIN7* polar localization with auxin gradients suggests an involvement of *PIN7* in this distribution. Therefore, we examined *DR5rev::GFP* expression in *pin7* early embryogenesis. Whereas wild-type embryos accumulated GFP in the apical cell lineage (see Fig. 1a–c), more than half of the *pin7* embryos failed to establish the apical–basal auxin gradient (Fig. 4l; Supplementary Table), similar to embryos treated with auxin-efflux inhibitors (see Fig. 2b). This demonstrates a role for *PIN7*-dependent efflux in mediating the initial auxin gradient in early embryogenesis.

By comparison with their respective wild-type parental lines (Fig. 4a), the stereotypical pattern of early embryogenesis was affected in all *pin7* mutant lines as well as in *PIN7* RNAi embryos

(Supplementary Table). Specification of the apical daughter cell of the zygote was compromised, as demonstrated by horizontal instead of vertical division (Fig. 4d, compare with Fig. 4a) as well as by uniform expression of the basal cell marker *PIN7-Ds-GUS* (Fig. 4i, compare with Fig. 3h). Occasionally, *pin7* embryos failed to establish the proembryo completely, resembling filamentous structures at later stages (Fig. 4b, c). In most cases, the defects were confined to the lower region of the proembryo. This was demonstrated morphologically (Fig. 4d) and also by *PIN1* and *PIN1::GUS* apical markers, which frequently failed to be expressed in the lower region (Fig. 4j, k, compare with Fig. 3a, b). In some cases, two proembryos developed on top of each other (Fig. 4d). The lower proembryo usually developed more slowly and often retained morphological characteristics of a suspensor, as if the boundary between apical and basal embryo structures was not clearly defined. These defects strongly resembled other auxin-related mutants, such as *mp*, *bdl* and *gn*³¹ (Fig. 4m–o). Nonetheless, a detailed comparison of embryo phenotypes in these mutants revealed that preglobular defects in

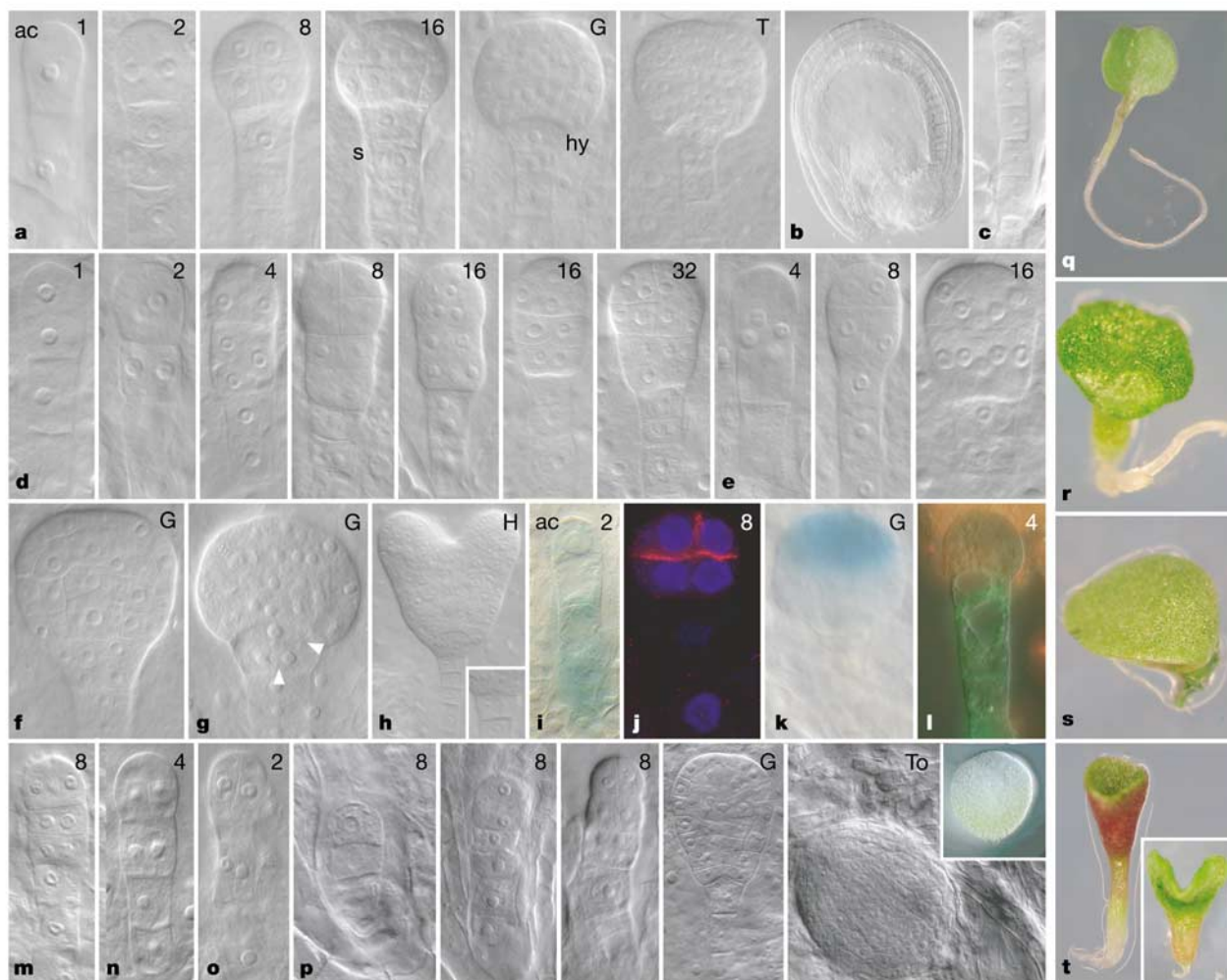


Figure 4 Abnormal embryogenesis in auxin transport and response mutants. **a**, Wild-type development. **b–e**, *pin7* mutants. Filamentous embryos with no apical cell established (**b**, **c**). Apical–basal boundary ill-defined (**d**), weaker defects in apical parts (**e**). **f–h**, Recovery of *pin7* proembryo after the globular stage (**f**), mild basal defects. Enlarged root pole with premature hypophysis division (**g**, arrowheads), abnormal cell divisions in the suspensor (**h**, inset). **i–l**, Abnormal marker expression in *pin7*. *PIN7-Ds-GUS* misexpressed in the apical cell (**ac**) (**i**), reduced apical domains of *PIN1* (**j**) and *PIN1::GUS* (**k**), *DR5rev::GFP* ectopic expression in suspensor (**l**). **m–o**, Early defects in *mp* (**m**), *bdl* (**n**) and *gn* (**o**) mutants. **p**, Compromised apical–basal axis in *pin1 pin3 pin4 pin7*

embryos; inset, *gn* ball-shaped embryo. **q–t**, *pin* multiple mutant seedling phenotypes. Cotyledon defects in *pin4 pin7* (**q**), apical defects and short root in *pin1 pin3 pin4* (**r**), *pin1 pin3 pin4 pin7* stronger (**s**) and weaker (**t**) phenotypes; inset, *gn* seedling. GUS staining in blue (**i**, **k**). Protein signal in red and DAPI nuclear counterstain in blue (**j**). GFP fluorescence in green (**l**). For embryo stages, numbers indicate the developmental stage according to actual number of proembryo cells of the corresponding wild-type stage: G, globular; T, triangular; H, heart; To, torpedo. Apical cell (**ac**), suspensor (**s**) and hypophysis (**hy**) are marked.

pin7 embryos were stronger and more penetrant (Supplementary Table). At the globular stage, *mp*, *bdl* and *gn* showed fully penetrant defects in the specification of the hypophysis, probably being the cause of the postembryonic rootless phenotype in these mutants. By contrast, *pin7* started to recover at the globular stage (Fig. 4f), coinciding with the onset of PIN1 basal localization, PIN4 expression and reversal of the auxin gradient. From then on, developmental aberrations were confined to the basal part of the embryo (Supplementary Table). The region of the hypophysis was enlarged and premature cell divisions occurred (Fig. 4g), accompanied by aberrant cell divisions in the suspensor (Fig. 4h, inset).

Mild defects at the basal embryo pole were also observed in the *pin1* mutant. Whereas the hypophysis was established and divided in about half of the globular and in all triangular wild-type embryos, this was the case for only about one-quarter and two-thirds of the respective embryos from *pin1*/PIN1 plants (Supplementary Table). Thus, *pin7* and *pin1* mutants displayed defects in establishing the apical and basal embryo poles at the sites of auxin response, which were not within, but adjacent to, the expression domains of polarly localized PIN1 and PIN7.

Functional redundancy among PIN proteins

Despite strong defects in early embryos, the majority of *pin7* mutants recovered and produced fertile plants that showed only mild auxin-related defects (not shown). This recovery coincided with the onset of PIN1 and PIN4 polar localization, suggesting functional redundancy among PIN genes. Therefore, we constructed double, triple and quadruple combinations of *pin7* with *pin1*, *pin3* and *pin4* mutants to generate loss of function of all embryonically expressed PIN genes. *pin4 pin7* showed apical defects that persisted into the seedling stage, such as aberrant cotyledon number (Fig. 4q), which were found in none of the single mutants. Stronger defects were observed in triple mutants; for example, apical defects including fused cotyledons and a very short root in *pin1 pin3 pin4* seedlings (Fig. 4r). *pin1 pin3 pin4 pin7* quadruple mutants showed pronounced defects in proembryo establishment, forming shrunken, filamentous or multi-layered structures at early stages (Fig. 4p). Later on, quadruple mutants failed to recover, in contrast to *pin7*, and produced malformed globular embryos. The terminal phenotype of quadruple mutants was variable (Supplementary Table). Most embryos displayed misplaced or fused cotyledons, deletion of apical structures and pronounced root pole defects. Some embryos completely failed to establish apical–basal polarity and were ball-shaped (Fig. 4p). Depending on the ecotype background, quadruple mutants were either embryo lethal or developed into seedlings with severe apical defects and no or a non-functional root (Fig. 4s, t). The aberrations in quadruple embryos were similar to the defects in embryos after interfering with auxin homeostasis (see Fig. 2h, i) or those in *gn* embryos (Fig. 4p, t, insets). Furthermore, *gn* embryos and seedlings show the same range of phenotypes as *pin* quadruple mutants (Supplementary Table). Thus, analysis of multiple mutants demonstrated functional redundancy among PIN proteins and identified PIN-dependent auxin transport as an essential mechanism for the recovery of the apical–basal axis in *pin7* mutants at later stages of embryogenesis.

Discussion

The analysis of axis formation in genetically tractable animal models, such as *Drosophila* and *Caenorhabditis*, has been greatly facilitated by the accessibility of the freshly laid egg to experimental manipulation. By contrast, flowering-plant embryos develop deep inside maternal tissues, seriously limiting a comparable approach. Although a molecular analysis of *Arabidopsis* genes involved in early patterning suggested a link to the plant signalling molecule auxin, the spatial and temporal requirement for auxin in embryogenesis and its role in axis formation were not shown. We have succeeded

in visualizing auxin and its response in early embryogenesis. The *in vitro* culture of *Arabidopsis* embryos enabled us to combine mutant and marker analysis with experimental interference of auxin homeostasis. This approach, in conjunction with a molecular analysis of auxin transport during early embryogenesis, proved to be instrumental in elucidating the mechanism of axis formation in plants.

Our results show a dynamic distribution of auxin and its response, and suggest routes of auxin transport during early embryogenesis (Fig. 5). Immediately after division of the zygote, auxin accumulates in the smaller apical cell, auxin response is activated and the cell is specified as founder of the proembryo (Fig. 5a). Auxin is actively provided here from the adjacent basal cell by PIN7-dependent transport, as both chemical efflux inhibition and *pin7* mutations cause failure of the establishment of the apical–basal auxin gradient and lead to auxin accumulation in the basal cell. Moreover, PIN7 is polarly localized to the apical plasma membrane of the basal cell, where it is perfectly positioned to provide auxin to the adjacent apical cell. This apical–basal auxin distribution mediates correct specification of the apical cell, because this specification is compromised in *pin7* and even more strongly in quadruple *pin* mutants as demonstrated by aberrations in both cell behaviour and marker expression. Thus, an actively maintained PIN-dependent auxin gradient is required for the specification of the apical cell, which later becomes the founder of the proembryo and all apical structures of the plant.

At the globular stage (Fig. 5b), bioactive auxin production probably starts in the apical embryo region. In support of this, auxin response is activated ectopically in this region following chemical inhibition of auxin transport or in *gn* mutants. At the same stage, PIN1 basal localization is established in the provascular cells, suggesting downward transport towards the region of the future root pole. Simultaneously, the asymmetric localization of PIN7 is reversed within the basal cells, mediating auxin transport out of the embryo. Subsequently, PIN4 expression starts at the basal pole of the embryo, supporting the action of both PIN1 and PIN7 (Fig. 5b). The PIN7-dependent auxin transport in the suspensor operates at a lower rate than that mediated by PIN1 and PIN4. As a result, the auxin gradient reverses, displaying its new maximum in

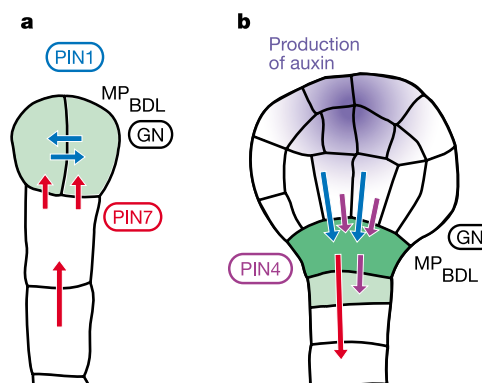


Figure 5 A model for a role of auxin in embryo patterning. Sites of auxin response and accumulation are shown in green. Arrows indicate routes of auxin efflux mediated by PIN1 (blue), PIN4 (purple) or PIN7 (red). Also depicted are proteins involved in embryo patterning and related to auxin transport (GN, encircled) or response (BDL, MP). **a**, Two-cell-stage embryo. Auxin accumulates in the proembryo through PIN7-dependent transport and triggers apical pole specification. **b**, Young globular embryo. Free auxin starts to be produced in the apical part (purple) and auxin transport routes reverse. Auxin accumulates in a PIN1- and PIN4-dependent manner in the hypophysis, triggering root pole specification.

the uppermost suspensor cell, which in response to auxin is specified to become the hypophysis—the founder of the future root meristem.

At this stage, PIN7 function becomes redundant, because *pin7* mutants recovered and were able to re-establish the axis. This recovery may be akin to *de novo* axis formation in other auxin-dependent processes, such as plant regeneration from callus or postembryonic organ initiation³². By contrast, *pin* quadruple mutants failed to recover at this stage and displayed strong *gn*-like defects. In the extreme case, these quadruple mutants were ball-shaped, entirely lacking apical–basal polarity, and the same effects were also caused by chemical inhibition of auxin transport. These findings indicate that the re-establishment of the axis at the globular stage is also mediated by a PIN-dependent auxin distribution, and that this mechanism is probably non-redundant. Thus, axis formation in embryogenesis appears to be a prime example of plant developmental plasticity, involving a self-repairing mechanism of auxin gradients.

In summary, our results provide the first coherent conceptual framework of how the apical–basal axis is established in plants. This process starts with the asymmetric cell division of the zygote, which generates auxin-transporting and auxin-responsive daughter cells. It is conceivable that this first step of axis initiation depends on the auxin transport from the maternal tissue into the zygote; however, owing to a lack of experimental data, this question remains open. Nonetheless, the elaboration of this initial axis is clearly auxin-dependent, because immediately an actively maintained auxin activity gradient is formed, which triggers first the specification of the apical pole and, after its reversal, that of the basal embryo pole. Thus, axis formation in plants involves active accumulation of a signalling molecule by efflux from polarized cells and a localized response. Active accumulation of a signalling molecule results in an inverse gradient with a maximum away from the source, in contrast to the known mechanisms of axis formation in other systems, which involve signalling molecules whose concentration steadily decreases from a maximum at the site of production^{33,34}. □

Methods

Materials

DR5rev::GFP consists of nine repeats of the auxin-response element (TGCTC) fused in inverse orientation to the CaMV minimal 35S promoter³⁵ and the ER-targeted eGFP coding sequence (Clontech). *GAL4/UAS* constructs for *DR5* expression of DTA were generated as described²⁷. Mutant lines of *gn*⁴ and *mp*⁴, *bdF*, *pin1* (ref. 20) and *pin4-3* (ref. 13) have been described. Full-length coding sequences of *PIN6* and *PIN7* (GenBank accession numbers AF087819 and AF087820), *PIN5* and *PIN8* were isolated from cDNA libraries prepared from seedlings, leaves and whole plants. *PIN7* RNAi plants expressed, from the CaMV 35S promoter, a construct spanning nucleotides 870–1,344 of *PIN7* cDNA sequence in both sense and antisense orientation, joined by the *uidA* coding sequence. *pin7-1*, -3 mutant lines with sequence-indexed insertions were identified in the Cold Spring Harbor gene-trap library (<http://genetrap.cshl.org/>) and *pin7-2*, *pin3-4*, *pin3-5* in the Signal Insertion Mutant Library (<http://signal.salk.edu/cgi-bin/tdnaexpress/>). *pin7-1*, -2 and -3 insertions were at positions 1,349, 1,836 and 2,174 from ATG, respectively. *pin3-4* and -5 insertions were at positions 1,014 and -51 from ATG, respectively. The *PIN7* enhancer-trap line from a library generated in our laboratory had the *Ds-GUS* transposon inserted at position 1,851 behind the *PIN7* stop codon. The *pin1 pin3 pin4 pin7* quadruple mutant was generated from *pin1*, *pin3-5*, *pin4-3* and *pin7-1* lines.

Growth conditions and microscopy

Plants were grown in a 16 h light/8 h dark cycle at 25/20 °C. For *in vitro* embryo culture, excised ovules were placed on X0.5 MS media containing 2% sucrose, 400 mg l⁻¹ glutamine and 0.3% Phytigel. For treatments, this medium was supplemented with 20–50 μM of NAA, 2,4-D or NPA, or 10–20 μM of BFA. Plates were kept in the dark at 22 °C for up to 7 days. At different times, embryos were excised from the ovules for microscopic analysis. For each condition and stage, at least 40 embryos were analysed. Schiff staining was performed as described³⁶. For all treatments, markers and mutant phenotype analyses, we performed control experiments in the sister lines and analysed a sufficient number of embryos (see Supplementary Information). Microscopy was done on a Zeiss Axiophot equipped with an Axiocam HR CCD camera using differential interference contrast optics or epifluorescence. For confocal laser scanning microscopy, a Leica TCS SP was used. Images were processed in Adobe Photoshop.

Expression and immunolocalization analyses

For GFP visualization, samples were mounted in 5% glycerol without fixation and

inspected. Histochemical GUS activity staining was performed using a modified indigogenic method²⁷. *PIN2::GUS*³⁷, *PIN3::GUS*²² and *PIN4::GUS*¹³ have been described. *PIN1::GUS* and *PIN6::GUS* constructs were generated by fusing a polymerase chain reaction (PCR)-amplified fragment (nucleotides -1,289 to -5; -1,794 to -1) with the *uidA* gene. The *PIN7::GUS* translational fusion was generated by fusing *uidA* gene to the carboxy terminus of the *PIN7* coding sequence. *In situ* hybridization was performed as described⁷ with probes corresponding to cDNA regions of *PIN1* (nucleotides 679–1,149), *PIN3* (999–1,449) and *PIN7* (608–1,456). Anti-PIN1 (ref. 38), anti-PIN4 (ref. 13) and anti-IAA²⁸ (Phytodetek, Agdia) antibodies have been described. Anti-PIN7 antibodies were raised against recombinant proteins corresponding to amino acids 204 to 486 of PIN7, and were affinity-purified as described³⁸. Immunolocalization was done as described^{8,37}. Anti-PIN4, anti-PIN1 and anti-PIN7 antibodies were diluted 1:500, 1:200 and 1:50, respectively. Secondary fluorescein isothiocyanate (FITC)- or CY3-conjugated goat anti-rabbit antibodies (Dianova) were diluted 1:200 or 1:600, respectively. IAA immunolocalization³⁹ in embryos was done after prefixation with 3% 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDAC, Sigma) for 1 h. Anti-IAA antibody and secondary alkaline phosphatase-conjugated anti-mouse antibody (Novagen) were diluted 1:500 and 1:1,000, respectively.

Received 24 June; accepted 19 September 2003; doi:10.1038/nature02085.

- St Johnston, D. & Nüsslein-Volhard, C. The origin of pattern and polarity in the *Drosophila* embryo. *Cell* **68**, 201–219 (1992).
- Lyczak, R., Gomes, J. & Bowerman, B. Heads or tails: cell polarity and axis formation in the early *Caenorhabditis elegans* embryo. *Dev. Cell* **3**, 157–166 (2002).
- Jürgens, G. Apical–basal pattern formation in *Arabidopsis* embryogenesis. *EMBO J.* **20**, 3609–3616 (2001).
- Mayer, U., Torres Ruiz, R. A., Berleth, T., Miséra, S. & Jürgens, G. Mutations affecting body organization in the *Arabidopsis* embryo. *Nature* **353**, 402–407 (1991).
- Hamann, T., Mayer, U. & Jürgens, G. The auxin-insensitive *bodenlos* mutation affects primary root formation and apical–basal patterning in the *Arabidopsis* embryo. *Development* **126**, 1387–1395 (1999).
- Hardtke, C. & Berleth, T. The *Arabidopsis* gene *MONOPTEROS* encodes a transcription factor mediating embryo axis formation and vascular development. *EMBO J.* **17**, 1405–1411 (1998).
- Hamann, T., Benkova, E., Bäurle, I., Kientz, M. & Jürgens, G. The *Arabidopsis* *BODENLOS* gene encodes an auxin response protein inhibiting *MONOPTEROS*-mediated embryo patterning. *Genes Dev.* **16**, 1610–1615 (2002).
- Steinmann, T. *et al.* Coordinated polar localization of auxin efflux carrier PIN1 by GNOM ARF-GEF. *Science* **286**, 316–318 (1999).
- Geldner, N. *et al.* The *Arabidopsis* GNOM ARF-GEF mediates endosomal recycling, auxin transport, and auxin-dependent plant growth. *Cell* **112**, 219–230 (2003).
- Liu, C., Xu, Z. & Chua, N. Auxin polar transport is essential for the establishment of bilateral symmetry during early plant embryogenesis. *Plant Cell* **5**, 621–630 (1993).
- Hadfi, K., Speth, V. & Neuhaus, G. Auxin-induced developmental patterns in *Brassica juncea* embryos. *Development* **125**, 879–887 (1998).
- Sabatini, S. *et al.* An auxin-dependent distal organizer of pattern and polarity in the *Arabidopsis* root. *Cell* **99**, 463–472 (1999).
- Friml, J. *et al.* AtPIN4 mediates sink-driven auxin gradients and root patterning in *Arabidopsis*. *Cell* **108**, 661–673 (2000).
- Fischer-Iglesias, C., Sundberg, B., Neuhaus, G. & Jones, A. Auxin distribution and transport during embryonic pattern formation in wheat. *Plant J.* **26**, 115–129 (2001).
- Ribnick, D., Cohen, J., Hu, W. & Cooke, T. An auxin surge following fertilization in carrots: a mechanism for regulating plant totipotency. *Planta* **214**, 505–509 (2002).
- Friml, J. Auxin transport—shaping the plant. *Curr. Opin. Plant Biol.* **6**, 7–12 (2003).
- Rubery, P. & Sheldrake, A. Carrier-mediated auxin transport. *Planta* **118**, 101–121 (1974).
- Raven, J. Transport of indoleacetic acid in plant cells in relation to pH and electrical potential gradients, and its significance for polar IAA transport. *New Phytol.* **74**, 163–172 (1975).
- Friml, J. & Palme, K. Polar auxin transport—old questions and new concepts? *Plant Mol. Biol.* **49**, 273–284 (2002).
- Okada, K., Ueda, J., Komaki, M., Bell, C. & Shimura, Y. Requirement of the auxin polar transport system in early stages of *Arabidopsis* floral bud formation. *Plant Cell* **3**, 677–684 (1991).
- Rashotte, A., Brady, S., Reed, R., Ante, S. & Muday, G. Basipetal auxin transport is required for gravitropism in roots of *Arabidopsis*. *Plant Physiol.* **122**, 481–490 (2000).
- Friml, J., Wisniewska, J., Benková, E., Mendgen, K. & Palme, K. Lateral relocation of auxin efflux regulator AtPIN3 mediates tropism in *Arabidopsis*. *Nature* **415**, 806–809 (2002).
- Ottenschläger, I. *et al.* Gravity-regulated differential auxin transport from columella to lateral root cap cells. *Proc. Natl Acad. Sci. USA* **100**, 2987–2991 (2003).
- Yamaizumi, M., Mekada, E., Uchida, T. & Okada, Y. One molecule of Diptheria Toxin fragment A introduced into a cell can kill the cell. *Cell* **15**, 245–250 (1978).
- Haseloff, J. GFP variants for multispectral imaging of living cells. *Methods Cell Biol.* **58**, 139–151 (1999).
- Weijers, D., Geldner, N., Offringa, R. & Jürgens, G. Early paternal gene activity in *Arabidopsis*. *Nature* **414**, 709–710 (2001).
- Weijers, D. *Hormonal Regulation of Pattern Formation in the Arabidopsis Embryo*. Thesis, Univ. Leiden (2002).
- Caruso, J., Pence, V. & Leverone, L. in *Plant Hormones: Physiology, Biochemistry and Molecular Biology* (ed. Davies, P.) 43–447 (Kluwer Academic, The Netherlands, 1995).
- Delbarre, A., Muller, P., Imhoff, V. & Guern, J. Comparison of mechanisms controlling uptake and accumulation of 2,4-dichlorophenoxy acetic acid, naphthalene-1-acetic acid, and indole-3-acetic acid in suspension-cultured tobacco cells. *Planta* **198**, 532–541 (1996).
- Sundaresan, V. *et al.* Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev.* **14**, 1797–1810 (1995).
- Mayer, U., Büttner, G. & Jürgens, G. Apical–basal pattern formation in the *Arabidopsis* embryo: Studies on the role of the *gnom* gene. *Development* **117**, 149–162 (1993).
- Steeves, T. & Sussex, I. *Patterns in Plant Development* (Cambridge Univ. Press, Cambridge, 1989).

33. Teleman, A., Strigini, M. & Cohen, S. Shaping morphogen gradients. *Cell* **105**, 559–562 (2001).
34. Tabata, T. Genetics of morphogen gradients. *Nature Rev. Genet.* **2**, 620–630 (2001).
35. Ulmasov, T., Murfett, J., Hagen, G. & Guilfoyle, T. Aux/IAA proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. *Plant Cell* **9**, 1963–1971 (1997).
36. Braselton, J., Wilkinson, M. & Clulow, S. Feulgen staining of intact plant tissues for confocal microscopy. *Biotechnol. Histochem.* **71**, 84–87 (1996).
37. Friml, J., Benkova, E., Mayer, U., Palme, K. & Muster, G. Automated whole mount localisation techniques for plant seedlings. *Plant J.* **34**, 115–124 (2003).
38. Gälweiler, L. *et al.* Regulation of polar auxin transport by AtPIN1 in *Arabidopsis* vascular tissue. *Science* **282**, 2226–2230 (1998).
39. Moctezuma, E. Changes in auxin patterns in developing gynophores of the peanut plant (*Arachis hypogaea* L.). *Ann. Bot.* **83**, 235–242 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Hiller, M. Kientz, G. Martin and M. L. O. Mendes for technical assistance, and C. Luschnig and K. Palme for providing material. Sequence-indexed *Arabidopsis* insertion mutants were obtained from the Salk Institute Genomic Analysis Laboratory and Cold Spring Harbor Laboratory. We are grateful to K. Cornelis and N. Geldner for critical reading of the manuscript. This work was supported by the Volkswagen Stiftung programme (M.S., J.F.), Landesgraduiertenförderung (A.V.), and the Research Council for Earth and Lifesciences (ALW), with financial aid from the Dutch Organization of Scientific Research (NWO) (D.W.) and the Deutsche Forschungsgemeinschaft.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.F. (jiri.friml@zmbp.uni-tuebingen.de).

A common origin for cosmic explosions inferred from calorimetry of GRB030329

E. Berger¹, S. R. Kulkarni¹, G. Pooley⁴, D. A. Frail⁵, V. McIntyre⁶, R. M. Wark⁷, R. Sari², A. M. Soderberg¹, D. W. Fox¹, S. Yost³ & P. A. Price⁸

¹Caltech Optical Observatories 105-24, ²Theoretical Astrophysics 130-33,

³Space Radiation Laboratory 220-47, California Institute of Technology, Pasadena, California 91125, USA

⁴Mullard Radio Astronomy Observatory, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, UK

⁵National Radio Astronomy Observatory, P O Box 0, Socorro, New Mexico 87801, USA

⁶Australia Telescope National Facility, CSIRO, P O Box 76, Epping, New South Wales 1710, Australia

⁷Australia Telescope National Facility, CSIRO, Locked Bag 194, Narrabri, New South Wales 2390 Australia

⁸RSAA, ANU, Mt Stromlo Observatory, via Cotter Rd, Weston Creek, Australian Capital Territory, 2611, Australia

Past studies^{1–3} have suggested that long-duration γ -ray bursts have a ‘standard’ energy of $E_\gamma \approx 10^{51}$ erg in the ultra-relativistic ejecta, after correcting for asymmetries in the explosion (‘jets’). But a group of sub-energetic bursts, including the peculiar GRB980425 associated⁴ with the supernova SN1998bw ($E_\gamma \approx 10^{48}$ erg), has recently been identified^{2,3}. Here we report radio observations of GRB030329 that allow us to undertake calorimetry of the explosion. Our data require a two-component explosion: a narrow (5° opening angle) ultra-relativistic component responsible for the γ -rays and early afterglow, and a wide, mildly relativistic component that produces the radio and optical afterglow more than 1.5 days after the explosion. The total energy release, which is dominated by the wide component, is similar^{1–3,5} to that of other γ -ray bursts, but the contribution of the γ -rays is energetically minor. Given the firm link^{6,7} of GRB030329 with SN2003dh, our result indicates a common origin for cosmic explosions in which, for reasons not yet understood, the energy in the highest-velocity ejecta is extremely variable.

We initiated observations of the nearby GRB030329 (redshift $z = 0.1685$) in the centimetre band approximately 13.8 h after the burst. The log of the observations and the resulting light curves are displayed in Tables 1 and 2 and Fig. 1. The afterglow was also observed extensively in the millimetre (100 GHz) and submillimetre (250 GHz) bands⁸. Although this is the brightest radio afterglow detected to date, the low redshift results in a peak luminosity, $L_{\nu,p}(8.5 \text{ GHz}) \approx 1.8 \times 10^{31} \text{ erg s}^{-1} \text{ Hz}^{-1}$, typical⁹ of other long-duration γ -ray bursts (GRBs).

The observed rapid decline, $F_\nu \propto t^{-1.9}$ at $t \geq 10$ d and the decrease in peak flux at $\nu \approx 22.5$ GHz (Fig. 1) are the hallmarks of a collimated explosion. In this framework¹⁰, the sharp decline (or ‘jet break’) occurs at the time, t_j , when $\Gamma(t_j) \approx \theta_j^{-1}$ owing to relativistic aberration (‘beaming’) and rapid sideways expansion; here Γ is the bulk Lorentz factor and θ_j is the opening angle of the jet. We model the afterglow emission (compare refs 5, 11) from 4.9 to 250 GHz assuming a uniform¹⁰ as well as a ‘wind’¹² (particle density profile, $\rho \propto r^{-2}$, where r is the distance from the source) circumburst medium. Neither model is strongly preferred, but $t_{j,\text{rad}} \approx 9.8$ d is required (Fig. 1).

Using the inferred particle density of $n \approx 1.8 \text{ cm}^{-3}$ and assuming a γ -ray efficiency, $\epsilon_\gamma = 0.2$ (see ref. 3) we infer $\theta_{j,\text{rad}} \approx 0.3$ rad, or 17° . The kinetic energy in the explosion corrected for collimation is $E_K = f_b E_{K,\text{iso}} \approx 2.5 \times 10^{50}$ erg, where $f_b = [1 - \cos(\theta_j)]$ is the beaming fraction and $E_{K,\text{iso}}$ is the isotropic equivalent kinetic energy. This

value is comparable to that inferred from modelling of other afterglows⁵.

In contrast to the above discussion, Price *et al.*¹³ note a sharp break in the optical afterglow at $t = 0.55$ d (Fig. 2). The X-ray flux¹⁴ tracks the optical afterglow for the first day, with a break consistent with that seen in the optical. Thus the break at 0.55 d is not due to a change in the ambient density as, for typical parameters^{15,16}, the X-ray emission is not sensitive to density. However, unlike the optical emission, the X-ray flux at later times continues to decrease monotonically. Thus, we conclude that there are two emitting components: one responsible for the early optical and X-ray emission, and the other responsible for the optical emission beyond 1.5 d.

The first component, given the characteristic t^{-2} decay for both the X-ray and optical emission, is reasonably modelled by a jet. For the parameters used above (n, ϵ_γ), the opening angle is 0.09 rad or 5° .

The resurgence in the optical emission at 1.5 d requires a second component. An increase in the ambient density cannot explain this resurgence as the predicted decrease in radio luminosity, arising from the increase in synchrotron self-absorption, is not observed (Fig 1). An increase in the energy of the first component, for example by successive shells with lower Lorentz factors as advocated by Granot *et al.*¹⁷, is ruled out by the lack⁸ of strong radio or millimetric emission expected¹⁸ from reverse shocks.

Thus, by a process of elimination, we are led to a two-component explosion model in which the first component (a narrow jet, 5°)

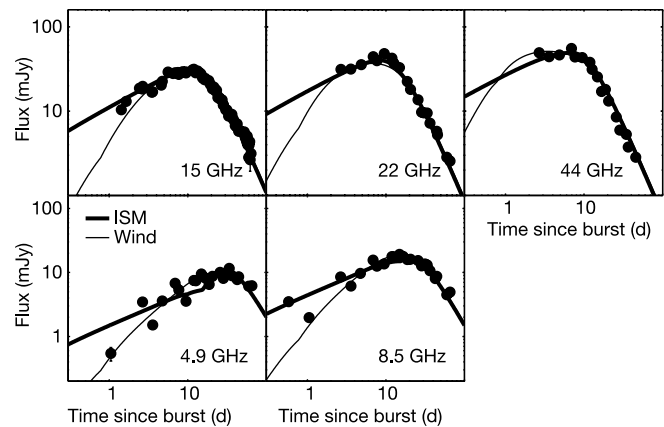


Figure 1 Radio light curves of the afterglow of GRB030329. All measurements include 1σ error bars, which in most cases are smaller than the symbols. The data are summarized in Tables 1 and 2. The solid lines are models of synchrotron emission from collimated relativistic ejecta expanding into uniform¹⁰ (thick) and wind¹² $\rho \propto r^{-2}$ (thin) circumburst media; these models include the millimetre and submillimetre data⁸. We find $\chi^2_\nu = 31.3$ and 39.8 (164 degrees of freedom) for the uniform density and wind models, respectively; these include a 2% systematic error added in quadrature to each measurement. The large values of χ^2_ν are dominated by interstellar scintillation (ISS) at $\nu \leq 15$ GHz and mild deviations from the expected smooth behaviour at the high frequencies. Comparing the data and models, we find r.m.s. flux modulations of 0.25 at 4.9 GHz, 0.15 at 8.5 GHz, and 0.08 at 15 GHz, as well as a drop by a factor of three in the level of modulation from ~ 3 to 40 d. These properties are expected in weak ISS as the fireball expands on the sky. The inferred source size of about $20 \mu\text{as}$ (that is, $\sim 2 \times 10^{17}$ cm) at $t \approx 15$ d is in close agreement with theoretical expectations²². In the uniform density model the jet break occurs at $t \approx 10$ d corresponding to an opening angle, $\theta_j \approx 0.3$ (17°). From the derived synchrotron parameters (at $t = t_j$: self-absorption frequency $\nu_a \approx 19$ GHz, synchrotron characteristic frequency $\nu_m \approx 43$ GHz, synchrotron peak flux $F_{\nu,m} \approx 96$ mJy) we find an isotropic kinetic energy, $E_{K,\text{iso}} \approx 5.6 \times 10^{51} \nu_{c,13}^{1/4}$ erg, a circumburst density $n = 1.8 \nu_{c,13}^{3/4} \text{ cm}^{-3}$, and the fractions of energy in the relativistic electrons and magnetic field of $0.16 \nu_{c,13}^{1/4}$ and $0.10 \nu_{c,13}^{-5/4}$, respectively; here $\nu_c = 10^{13} \nu_{c,13}$ is the synchrotron cooling frequency, and a constraint on inverse Compton cooling as advocated by ref. 23 indicates $\nu_{c,13} \leq 1$. The beaming-corrected kinetic energy is $E_K \approx 2.5 \times 10^{50} \nu_{c,13}^{1/4}$ erg, typical of other well-studied long-duration GRBs⁵. The parameters derived from the wind model are consistent with those from the uniform density model to within 10%.

Table 1 Radio observations made with the VLA and the ATCA

Epoch (UT)	Δt (d)	$F_{1.43}$ (mJy)	$F_{4.86}$ (mJy)	$F_{8.46}$ (mJy)	$F_{15.0}$ (mJy)	$F_{22.5}$ (mJy)	$F_{43.3}$ (mJy)
Mar.30.06	0.58	—	—	3.50 ± 0.06	—	—	—
Mar.30.53	1.05	—	0.54 ± 0.13	1.98 ± 0.17	—	—	—
Apr.1.13	2.65	< 0.21	3.45 ± 0.05	8.50 ± 0.05	19.68 ± 0.14	30.40 ± 0.06	46.63 ± 0.18
Apr.2.05	3.57	< 0.30	1.51 ± 0.05	6.11 ± 0.04	16.98 ± 0.19	31.59 ± 0.14	44.17 ± 0.35
Apr.3.21	4.76	< 0.36	3.58 ± 0.04	9.68 ± 0.03	22.59 ± 0.12	35.57 ± 0.09	46.32 ± 0.23
Apr.5.37	6.89	< 0.40	6.77 ± 0.08	15.56 ± 0.06	28.58 ± 0.20	44.09 ± 0.15	55.33 ± 0.43
Apr.6.16	7.68	< 0.25	5.34 ± 0.10	12.55 ± 0.21	27.26 ± 0.21	39.68 ± 0.20	43.81 ± 1.00
Apr.7.97	9.49	< 0.68	3.55 ± 0.11	13.58 ± 0.09	28.50 ± 0.23	48.16 ± 0.23	43.06 ± 1.33
Apr.10.38	11.90	< 0.58	7.51 ± 0.08	17.70 ± 0.05	31.40 ± 0.25	42.50 ± 0.14	37.86 ± 0.46
Apr.11.17	12.69	—	7.42 ± 0.09	17.28 ± 0.10	29.60 ± 0.29	36.84 ± 0.16	31.26 ± 0.51
Apr.13.35	14.87	—	9.49 ± 0.13	19.15 ± 0.08	26.78 ± 0.33	32.69 ± 0.13	25.44 ± 0.51
Apr.15.14	16.66	—	8.21 ± 0.08	17.77 ± 0.10	24.50 ± 0.31	—	17.10 ± 0.71
Apr.17.20	18.72	< 0.63	6.50 ± 0.11	15.92 ± 0.07	22.02 ± 0.25	22.41 ± 0.08	18.07 ± 0.28
Apr.19.06	20.58	—	8.66 ± 0.10	16.08 ± 0.06	18.35 ± 0.24	18.03 ± 0.11	13.15 ± 0.29
Apr.24.18	25.70	—	10.04 ± 0.08	15.34 ± 0.06	13.93 ± 0.26	13.63 ± 0.13	8.54 ± 0.48
Apr.26.92	28.44	< 0.58	8.05 ± 0.08	12.67 ± 0.09	11.82 ± 0.26	9.75 ± 0.23	5.95 ± 0.62
Apr.28.96	30.48	—	—	—	10.40 ± 0.33	9.53 ± 0.21	—
Apr.29.99	31.51	—	9.80 ± 0.09	13.55 ± 0.07	—	—	—
May 2.06	33.58	—	11.62 ± 0.08	13.10 ± 0.06	—	9.52 ± 0.14	—
May 3.07	34.59	—	—	—	—	—	5.30 ± 0.32
May 5.00	36.52	—	8.90 ± 0.08	10.64 ± 0.06	8.58 ± 0.17	7.20 ± 0.09	3.75 ± 0.26
May 11.03	42.55	—	7.72 ± 0.13	8.04 ± 0.08	7.03 ± 0.19	—	—
May 13.03	44.55	—	8.57 ± 0.09	8.68 ± 0.08	5.77 ± 0.22	5.75 ± 0.10	—
May 14.00	45.52	—	—	—	—	5.23 ± 0.17	2.84 ± 0.23
May 28.03	59.55	—	6.08 ± 0.10	4.48 ± 0.09	2.82 ± 0.21	2.84 ± 0.20	—
June 4.01	66.53	1.94 ± 0.06	6.20 ± 0.08	4.93 ± 0.06	—	2.56 ± 0.12	—

Observations commenced on March 30.06 UT, with a single 7-h observation with the Australia Telescope Compact Array (ATCA) on March 30.53 UT. In the initial observation we detected a point source at $\alpha(J2000) = 10^{\text{h}} 44^{\text{m}} 49.95^{\text{s}}$, $\delta(J2000) = 21^{\circ} 31' 17.38''$, with an uncertainty of about 0.1 arcsec in each coordinate, consistent with the position of the optical counterpart. In all Very Large Array (VLA) observations we used the standard continuum mode with 2×50 MHz bands. At 22.5 and 43.3 GHz we used referenced pointing scans to correct for the systematic 10–20 arcsec pointing errors of the VLA antennas. We used the extra-galactic sources 3C147 (J0542+498) and 3C286 (J1331+305) for flux calibration, while the phase was monitored using J1111+199 at 1.43 GHz and J1051+213 at all other frequencies. The ATCA observations were performed at 4.80, 6.21, 8.26 and 9.02 GHz with a bandwidth of 64 MHz in each frequency. The phase was monitored using J1049+215, while the flux was calibrated using J1934–638. The data were reduced and analysed using the Astronomical Image Processing System (VLA) and the Multichannel Image Reconstruction, Image Analysis and Display package (ATCA). The flux density and uncertainty were measured from the resulting maps by fitting a gaussian model to the afterglow. In addition to the r.m.s. noise in each measurement we estimate a systematic uncertainty of about 2% due to uncertainty in the absolute flux calibration.

with initially larger Γ is responsible for the γ -ray burst and the early optical and X-ray afterglow including the break at 0.55 d, while the second component (a wider jet, 17°) powers the radio afterglow and late optical emission (Fig 2). The break due to the second component is readily seen in the radio afterglow, but is masked by SN2003dh in the optical bands, thus requiring careful subtraction (Fig 2). Such a two-component jet finds a natural explanation in the collapsar model¹⁹.

The beaming-corrected γ -ray energy, emitted by the narrow jet, is only $E_\gamma \approx 5 \times 10^{49}$ erg, significantly lower than the strong clustering³ around 1.3×10^{51} erg seen in most bursts. Similarly, the beaming-corrected X-ray luminosity¹⁴ at $t = 10$ h, a proxy for the kinetic energy of the afterglow on that timescale, is $L_{X,10} \approx 3 \times 10^{43}$ erg s⁻¹, a factor of ten below the tightly clustered values² for most other bursts. However, the second component, which is mildly relativistic

(as determined by the lower energy peak of its spectrum), carries the bulk of the energy, as indicated by our modelling of the radio emission. We note that our model, with the energy in the lower Lorentz factor component dominating over the narrow ultra-relativistic component, is not consistent with the “universal standard jet” model²⁰.

The afterglow calorimetry presented here has important ramifications for our understanding of GRB engines. Recently, we have come to recognize a subclass of cosmological GRBs marked by rapidly fading afterglows at early time (that is, similar to GRB030329). These events are sub-energetic^{2,3} in E_γ and early X-ray afterglow luminosity. However, as demonstrated by our calorimetry of GRB030329, such bursts may have total explosive yields similar to other GRBs (Fig 3).

This leads to the following conclusions. First, radio calorimetry,

Table 2 Radio observations made with the Ryle Telescope at Cambridge, UK

Epoch (UT)	Δt (d)	$F_{15.3}$ (mJy)	Epoch (UT)	Δt (d)	$F_{15.3}$ (mJy)
Mar. 30.91	1.43	10.38 ± 0.28	Apr. 21.72	23.24	17.63 ± 0.29
Mar. 31.12	1.64	13.05 ± 0.28	Apr. 22.66	24.18	14.51 ± 0.49
Mar. 31.91	2.43	18.66 ± 0.28	Apr. 23.33	24.85	14.62 ± 0.49
Apr. 1.12	2.64	18.29 ± 0.28	Apr. 25.81	27.33	13.60 ± 0.65
Apr. 1.98	3.50	16.75 ± 0.27	Apr. 26.82	28.34	11.78 ± 0.52
Apr. 3.07	4.59	20.36 ± 0.45	Apr. 29.82	31.34	10.35 ± 0.49
Apr. 4.09	5.61	29.13 ± 0.52	May 1.63	33.15	8.73 ± 0.52
Apr. 4.97	6.49	27.97 ± 0.26	May 4.80	36.32	9.15 ± 0.50
Apr. 5.97	7.49	28.69 ± 0.26	May 6.83	38.35	7.87 ± 0.50
Apr. 7.06	8.58	29.29 ± 0.49	May 8.73	40.25	6.70 ± 0.50
Apr. 7.89	9.41	29.15 ± 0.44	May 10.76	42.28	6.49 ± 0.50
Apr. 9.89	11.41	30.78 ± 0.51	May 15.76	47.28	5.74 ± 0.50
Apr. 11.05	12.57	28.52 ± 0.51	May 20.70	52.22	5.69 ± 0.53
Apr. 11.88	13.40	29.92 ± 0.44	May 22.76	54.28	4.78 ± 0.78
Apr. 13.05	14.57	27.90 ± 0.44	May 24.76	56.28	4.31 ± 0.55
Apr. 13.87	15.39	24.74 ± 0.44	May 25.56	57.08	5.04 ± 0.84
Apr. 14.82	16.34	23.60 ± 0.32	May 26.75	58.27	3.99 ± 0.63
Apr. 16.96	18.48	23.06 ± 0.24	May 28.76	60.28	3.96 ± 0.58
Apr. 17.92	19.44	20.51 ± 0.24	May 29.82	61.34	4.35 ± 0.50
Apr. 19.95	21.47	19.27 ± 0.38	May 30.76	62.28	2.65 ± 0.72
Apr. 20.72	22.24	17.53 ± 0.33	June 2.54	64.06	3.13 ± 0.76

All observations were made at 15.3 GHz by interleaving 15-min scans of GRB030329 with 2.5-min scans of the phase calibrator J1051+2119. The absolute flux scale was calibrated using 3C48 and 3C286. We used five antennas providing ten baselines in the range 35–140 m. As the position of the source is well known, the in-phase component of the vector sum of the ten base lines was used as an unbiased estimate of the flux density. The typical r.m.s. fluctuation on the signal in a 32-s integration period is 6 mJy. We also add a systematic uncertainty of ~2% due to uncertainty in the absolute flux calibration.

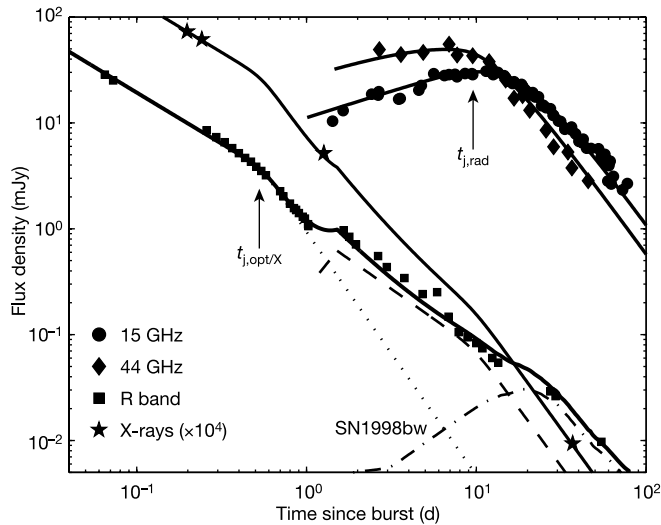


Figure 2 Radio to X-ray light curves of the afterglow of GRB030329. The optical data, from Price *et al.*¹³ and the GRB Coordinates Network^{24–26}, have been corrected for Galactic extinction, $A_R = 0.067$ mag; we note that the latter are preliminary data. The dotted line is the model proposed by Price *et al.*¹³ for the early optical emission, with $t_{j,opt} \approx 0.55$ d. The dashed line is an extrapolation of the uniform density model presented in Fig. 1 to the optical R -band with $\nu_{c,13} = 2$; this value is somewhat larger than the rough limit discussed in Fig. 1 legend but may be consistent with the uncertainty in the model parameters. The model in the X-ray band is based on the measured¹⁴ optical to X-ray spectral slope and an extrapolation of the uniform density model presented in Fig. 1. The sharp increase in the optical flux at $t \lesssim 1.5$ d is due to the deceleration of the slower second jet component. Finally, the dot-dashed line is the optical emission from SN1998bw at the redshift of GRB030329, $z = 0.1685$, used as a proxy for SN2003dh⁶. The solid line in the R -band is a combination of the supernova and the two jet components, whereas in the radio and millimetre bands it is the uniform density model presented in Fig. 1. In the X-ray band the model is dominated by the narrow jet component. Although this two-component jet model provides a reasonable fit to the data, there are still some discrepancies which could be resolved by accurate photometry and a more careful subtraction of SN2003dh. The latter would also allow a more precise determination of the putative second jet break in the optical band at $t_{j,rad}$.

which is sensitive to all ejecta with Γ in excess of a few times unity, shows that the explosive yield of the nearest ‘classical’ event, GRB030329, is dominated by mildly relativistic ejecta. Ultra-relativistic ejecta, which produced the γ -ray emission, are energetically unimportant. Second, the total energy yield of GRB030329 is similar to those estimated for other bursts. Along these lines, the enigmatic GRB980425 associated⁴ with the nearby supernova SN1998bw also has negligible γ -ray emission, $E_{\gamma,iso} \approx 8 \times 10^{47}$ erg; however, radio calorimetry²¹ shows that even this extreme event had a similar explosive energy yield (Fig. 3). The newly recognized class of cosmic explosions, the X-ray flashes (J. Heire, J.J.M. in’t Zand and S.R.K., manuscript in preparation), exhibit little or no γ -ray emission but appear to have comparable X-ray and radio afterglows to those of GRBs. Thus, the commonality of the total energy yield indicates a common origin, but apparently the ultra-relativistic output is highly variable. Unravelling what physical parameter is responsible for the variation in the ‘purity’ (ultra-relativistic output) of the engine appears to be the next frontier in the field of cosmic explosions. $\square$

Received 6 June; accepted 14 August 2003; doi:10.1038/nature01998.

1. Frail, D. A. *et al.* Beaming in gamma-ray bursts: Evidence for a standard energy reservoir. *Astrophys. J.* **562**, L55–L58 (2001).
2. Berger, E., Kulkarni, S. R. & Frail, D. A. A standard kinetic energy reservoir in gamma-ray burst afterglows. *Astrophys. J.* **590**, 379–385 (2003).
3. Bloom, J. S., Frail, D. A., Kulkarni, S. R. GRB energetics and the GRB Hubble diagram: Promises and limitations. *Astrophys. J.* **594**, 674–683 (2003).
4. Galama, T. J. *et al.* An unusual supernova in the error box of the gamma-ray burst of 25 April 1998. *Nature* **395**, 670–672 (1998).
5. Panaitescu, A. & Kumar, P. Properties of relativistic jets in gamma-ray burst afterglows. *Astrophys. J.*

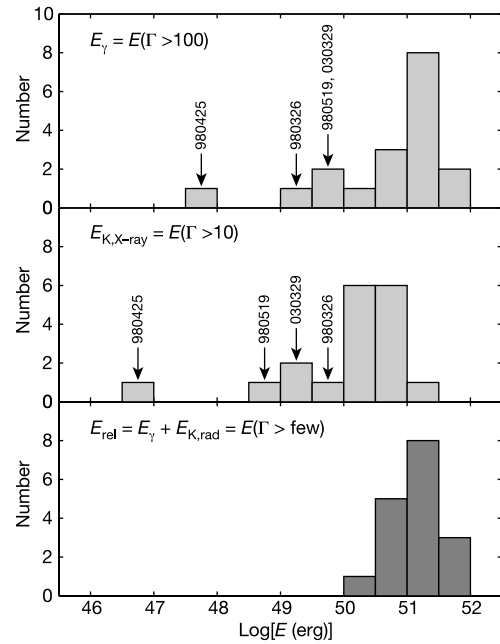


Figure 3 Histograms of various energies measured for GRBs. We plot the beaming-corrected γ -ray energy³, E_γ , the kinetic energy inferred from X-rays at $t = 10$ h (ref. 2), $E_{K,X}$, and total relativistic energy, $E_{rel} = E_\gamma + E_K$, where E_K is the beaming-corrected kinetic energy inferred^{5,21} from the broad-band afterglow. The energy in X-rays, $E_{K,X} = L_X t / \epsilon_e (\alpha_X - 1)$, with $t = 10$ h, $\epsilon_e = 0.1$, and $\alpha_X = 1.3$ is the median decay rate in the X-ray band. For GRB980519 we find that the evolution of the radio emission requires a much wider jet, $\theta_j \approx 0.3$, than what is inferred from the optical, $\theta_j \approx 0.05$; here we assume $z = 1$. We therefore infer $E_K \approx 2 \times 10^{50}$ erg from the radio data compared to $E_\gamma \approx 4 \times 10^{49}$ erg. The γ -ray energy of GRB980425 is an upper limit as the degree of collimation is not known. For the kinetic energy we use the value derived by Li and Chevalier²¹ on the basis of the radio evolution of SN1998bw. There is a significantly wider dispersion E_γ and $E_{K,X}$ as compared to the total explosive yield. This indicates that engines in cosmic explosions produce approximately the same quantity of energy, thus pointing to a common origin, but the quality of these engines, as indicated by ultra-relativistic output, varies widely.

6. Stanek, K. Z. *et al.* Spectroscopic discovery of the supernova 2003dh associated with GRB 030329. *Astrophys. J.* **591**, L17–L20 (2003).
7. Hjorth, J. *et al.* A very energetic supernova associated with the γ -ray burst of 29 March 2003. *Nature* **423**, 847–850 (2003).
8. Sheth, K. *et al.* GRB 030329 at millimeter wavelengths: OVRO, BIMA, & MAMBO observations. *Astrophys. J.* (submitted).
9. Frail, D. A., Kulkarni, S. R., Berger, E. & Wieringa, M. H. A complete catalog of radio afterglows: The first five years. *Astron. J.* **125**, 2299–2306 (2003).
10. Sari, R., Piran, T. & Halpern, J. P. Jets in gamma-ray bursts. *Astrophys. J.* **519**, L17–L20 (1999).
11. Berger, E. *et al.* A jet model for the afterglow emission from GRB 000301C. *Astrophys. J.* **545**, 56–62 (2000).
12. Chevalier, R. A. & Li, Z. Wind interaction models for gamma-ray burst afterglows: The case for two types of progenitors. *Astrophys. J.* **536**, 195–212 (2000).
13. Price, P. A. *et al.* The bright optical afterglow of the nearby γ -ray burst of 29 March 2003. *Nature* **423**, 844–847 (2003).
14. Tiengo, A. *et al.* The X-ray afterglow of GRB030329. Preprint at (<http://arXiv.org/astro-ph/0305564>) (2003).
15. Kumar, P. The distribution of burst energy and shock parameters for gamma-ray bursts. *Astrophys. J.* **538**, L125–L128 (2000).
16. Freedman, D. L. & Waxman, E. On the energy of gamma-ray bursts. *Astrophys. J.* **547**, 922–928 (2001).
17. Granot, J., Nakar, E., Piran, T. The variable light curve of GRB 030329. The case for refreshed shocks. Preprint at (<http://arXiv.org/astro-ph/0304563>) (2003).
18. Sari, R. & Mészáros, P. Impulsive and varying injection in gamma-ray burst afterglows. *Astrophys. J.* **535**, L33–L37 (2000).
19. MacFadyen, A. I., Woosley, S. E. & Heger, A. Supernovae, jets, and collapsars. *Astrophys. J.* **550**, 410–425 (2001).
20. Rossi, E., Lazzati, D. & Rees, M. J. Afterglow light curves, viewing angle and the jet structure of γ -ray bursts. *Mon. Not. R. Astron. Soc.* **332**, 945–950 (2002).
21. Li, Z. & Chevalier, R. A. Radio supernova SN 1998BW and its relation to GRB 980425. *Astrophys. J.* **526**, 716–726 (1999).
22. Galama, T. J. *et al.* Continued radio monitoring of the gamma-ray burst 991208. *Astrophys. J.* **585**, 899–907 (2003).
23. Sari, R. & Esin, A. A. On the synchrotron self-Compton emission from relativistic shocks and its

implications for gamma-ray burst afterglows. *Astrophys. J.* **548**, 787–799.

24. Henden, A., Canzian, B., Zeh, A. & Klose, S. GRB 030329: light curve flattens. *GRB Circ. Netw.* 2123 (2003).
25. Ibrahimov, M. A. et al. GRB 030329: BVRI photometry. *GRB Circ. Netw.* 2191 (2003).
26. Testa, V. et al. GRB 030329: VR photometry at TNG. *GRB Circ. Netw.* 2141 (2003).

Acknowledgements GRB research at Caltech is supported in part by NSF and NASA. We are indebted to S. Barthelmy and the GCN. The VLA is operated by the National Radio Astronomy Observatory, a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. The Australia Telescope is funded by the Commonwealth of Australia for operations as a National Facility managed by CSIRO. The Ryle Telescope is supported by PPARC.

Competing interests statement The authors declare that they have no competing interests.

Correspondence and requests for materials should be addressed to E.B. (ejb@astro.caltech.edu).

Evolution of the polarization of the optical afterglow of the γ -ray burst GRB030329

Jochen Greiner¹, Sylvio Klose², Klaus Reinsch³, Hans Martin Schmid⁴, Re'em Sari⁵, Dieter H. Hartmann⁶, Chryssa Kouveliotou⁷, Arne Rau¹, Eliana Palazzi⁸, Christian Straubmeier⁹, Bringfried Stecklum², Sergej Zharikov¹⁰, Gagrik Tovmassian¹⁰, Otto Bärnbantner¹¹, Christoph Ries¹¹, Emmanuel Jehin¹², Arne Henden¹³, Anlaug A. Kaas¹⁴, Tommy Grav¹⁵, Jens Hjorth¹⁶, Holger Pedersen¹⁶, Ralph A. M. J. Wijers¹⁷, Andreas Kaufer¹², Hye-Sook Park¹⁸, Grant Williams¹⁹ & Olaf Reimer²⁰

¹Max-Planck-Institut für extraterrestrische Physik, 85741 Garching, Germany

²Thüringer Landessternwarte Tautenburg, 07778 Tautenburg, Germany

³Universitäts-Sternwarte Göttingen, 37083 Göttingen, Germany

⁴Institut für Astronomie, ETH Zürich, 8092 Zürich, Switzerland

⁵California Institute of Technology, Theoretical Astrophysics 130-33, Pasadena, California 91125, USA

⁶Clemson University, Department of Physics and Astronomy, Clemson, South Carolina 29634, USA

⁷NSSTC, SD-50, 320 Sparkman Drive, Huntsville, Alabama 35805, USA

⁸Istituto di Astrofisica Spaziale e Fisica Cosmica, CNR, Sezione di Bologna, 40129 Bologna, Italy

⁹Physikalisches Institut, Universität Köln, 50937 Köln, Germany

¹⁰Instituto de Astronomia, UNAM, 22860 Ensenada, Mexico

¹¹Wendelstein-Observatorium, Universitätssternwarte, 81679 München, Germany

¹²European Southern Observatory, Alonso de Cordova 3107, Vitacura, Casilla 19001, Santiago 19, Chile

¹³Universities Space Research Association, US Naval Observatory, PO Box 1149, Flagstaff, Arizona 86002, USA

¹⁴Nordic Optical Telescope, 38700 Santa Cruz de La Palma, Spain

¹⁵University of Oslo, Institute for Theoretical Astrophysics, 0315 Oslo, Norway, and Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA

¹⁶Astronomical Observatory, NBI/AFG, University of Copenhagen, 2100 Copenhagen Ø, Denmark

¹⁷Astronomical Institute Anton Pannekoek, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

¹⁸Lawrence Livermore National Laboratory, University of California, PO Box 808, Livermore, California 94551, USA

¹⁹MMT Observatory, University of Arizona, Tucson, Arizona 85721, USA

²⁰Theoretische Weltraum- und Astrophysik, Ruhr-Universität Bochum, 44780 Bochum, Germany

The association of a supernova with GRB030329^{1,2} strongly supports the ‘collapsar’ model³ of γ -ray bursts, where a relativistic jet⁴ forms after the progenitor star collapses. Such jets cannot be spatially resolved because γ -ray bursts lie at cosmological distances; their existence is instead inferred from ‘breaks’ in the light curves of the afterglows, and from the theoretical desire to

reduce the estimated total energy of the burst by proposing that most of it comes out in narrow beams. Temporal evolution of the polarization of the afterglows^{5–7} may provide independent evidence for the jet structure of the relativistic outflow. Small-level polarization (~ 1 –3 per cent)^{8–17} has been reported for a few bursts, but its temporal evolution has yet to be established. Here we report polarimetric observations of the afterglow of GRB030329. We establish the polarization light curve, detect sustained polarization at the per cent level, and find significant variability. The data imply that the afterglow magnetic field has a small coherence length and is mostly random, probably generated by turbulence, in contrast with the picture arising from the high polarization detected in the prompt γ -rays from GRB021206 (ref. 18).

GRB030329 triggered the High Energy Transient Explorer, HETE-II, on 29 March 2003 (11:37:14.67 UT)¹⁹. The discovery of the burst optical afterglow^{20,21} was quickly followed by a redshift measurement²² for the γ -ray burster of $z = 0.1685$ (~ 800 Mpc), thus making GRB030329 the second-closest long-duration²³ γ -ray burst ever studied, after GRB980425 (ref. 3). The proximity of GRB030329 resulted in very bright prompt and afterglow emission, leading to the best-sampled afterglow to date. Detailed optical spectroscopy revealed an underlying supernova (SN2003dh)^{1,2} with an astonishing spectral similarity to SN1998bw (at

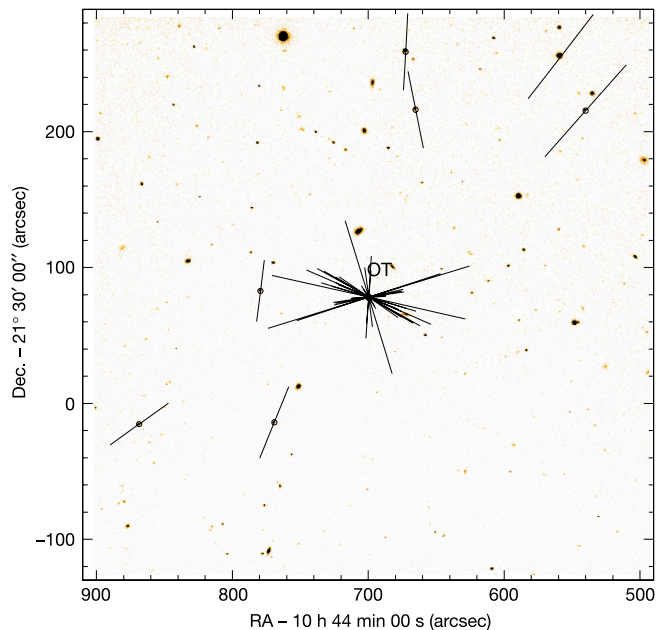


Figure 1 R band image of the field centred on the optical afterglow of GRB030329. The 27 FORS1 (Focal Reducer and Low Dispersion Spectrograph at the Very Large Telescope (VLT)/Antu) measurements are shown as ‘vectors’ where the length is a measure of the polarization degree, and the orientation indicates the position angle. While the afterglow varies in degree as well as angle, the polarization of seven field stars is constant within 0.1% in polarization degree and 1.5 degree in position angle throughout the 38 days. Linear polarization was measured from sets of exposures with different retarder-plate position angles. Imaging polarimetry was obtained during the first four nights (when the afterglow was brighter than 17th magnitude) from sixteen different retarder-plate angles, and from eight angles thereafter. Because the FORS1 polarization optics allows determination of the degree of polarization to an accuracy of $< 3 \times 10^{-4}$ and of the polarization angle to $\sim 0.2^\circ$, we consider the above variance of the field stars to represent the systematic error over the 38-day time period. Observations of polarimetric standard stars reproduced their tabulated values within 5%. The FORS1 retarder-plate zero-point angle of -1.2° was subtracted from the polarization angle. The position angle has a systematic uncertainty of $\pm 1.5^\circ$, which was added in quadrature to the statistical errors. OT, optical transient.

$z = 0.0085$, associated with GRB980425 (ref. 3)), thus strongly supporting the link of long-duration γ -ray bursts with core-collapse supernovae.

The apparent brightness of GRB030329 also afforded a unique opportunity to study the temporal evolution of polarization in the afterglow phase. Previous single measurements found low-level (1–3%) polarization^{8–15} in optical afterglows, and the only reports on variable polarization^{16,17} were based on few measurements with different instruments and modest signal-to-noise ratios. We have overcome the previous sampling limitations with 31 polarimetric observations of the afterglow of GRB030329 obtained with the same instrumentation (plus a few more with different instruments) over a time period of 38 days for a total of an unprecedented ~ 50 h of observations with an 8-m telescope. Here, for the first time, we establish a polarization light curve for an optical γ -ray burst afterglow.

We performed relative photometry, and from each pair of simultaneous measurements at orthogonal angles we derived the Stokes parameters U and Q . To obtain the intrinsic polarization of the γ -ray burst afterglow, we have to correct for Galactic interstellar polarization due to dust. We performed imaging polarimetry to derive the polarization parameters of seven stars in the field of GRB030329, and obtained an interstellar (dust) polarization correction of 0.45% at position angle 155° . Subtraction of the mean foreground polarization was performed in the Q – U plane ($Q_{\text{ip}} = 0.0027 \pm 0.0013$, $U_{\text{ip}} = -0.0033 \pm 0.0017$). Figure 1 shows an R-band image of the GRB030329 field with the polarization ‘vectors’ superimposed.

The temporal evolution of the degree and angle of polarization together with the R-band photometry is shown in Fig. 2. This figure demonstrates the presence of non-zero polarization, $\Pi \approx 0.3$ –2.5% throughout a 38-day period, with significant variability in degree and angle on timescales down to hours. Further, the spectropolarimetric data of the first three nights as well as the simultaneous R-

and K-band imaging polarimetry during the second night show that the relative polarization and the position angle are wavelength-independent (within the measurement errors of about 0.1%) over the entire spectral range. These data imply that polarization due to dust in the host galaxy of GRB030329 does not exceed $\sim 0.3\%$. Dust destruction in the vicinity of the γ -ray burst due to the strong radiation field would probably result in a monotonically decreasing polarization degree, which is not observed during the first days. We therefore conclude that the bulk of the observed variability is intrinsic to the afterglow.

Figure 2 shows that while the polarization properties show substantial variability (for which no simple empirical relationship is apparent), the R-band flux is a sequence of power laws. During each of the power-law decay phases the polarization is of the order of a few per cent, different from phase to phase, and variable within the phase, but not in tandem with the ‘bumps’ and ‘wiggles’ in the light curve. We observe a decreasing polarization degree shortly after the light curve break at ~ 0.4 days (as determined from optical^{21,24} and X-ray data^{25,26}). Rapid variations of polarization occur ~ 1.5 days after the burst, and could be related to the end of the transition period towards a new power-law phase starting at ~ 1.7 days. Polarization eventually rises to a level of $\sim 2\%$, which remains roughly constant for another two weeks. Late in the observation campaign the underlying type Ic SN2003dh^{1,2} increasingly contributes to the total light and probably also to the observed polarization properties. Asymmetries in this type of supernova can produce polarization of the order of 1% (refs 27, 28), as we observe towards the end of our campaign.

GRB030329 belongs to a growing group of bursts for which densely monitored afterglow light curves show significant ‘bumps’ and ‘wiggles’ relative to a simple power-law decay (most notably GRB021004 and GRB011211). These bumps and wiggles complicate the interpretation of the polarization properties. While the rapid decrease in polarization degree during the first night is consistent

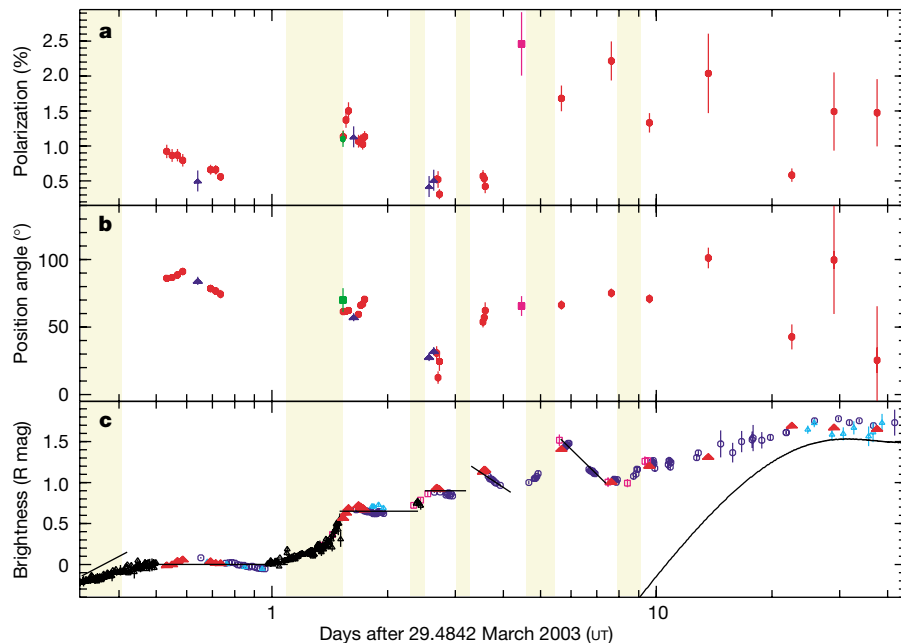


Figure 2 Evolution of the linear polarization during the first 38 days. **a, b**, The polarization degree in per cent and the position angle in degrees. The red data points are from imaging polarimetry with FORS1/VLT. Spectropolarimetry (blue symbols, 700–800 nm range) was performed during the first three nights. The green points were obtained with CAFOS at the 2.2-m telescope at Calar Alto on 30/31 March. The magenta data point was obtained with AFOSC at the 2.56-m NOT telescope on 2 April 2003. **c**, The residual R-band light curve after subtraction of the contribution of a power law $t^{-1.64}$ describing the undisturbed

decay during the time interval 0.5–1.2 days after the γ -ray burst (that is, after the early break at 0.4 days), thus leading to a horizontal curve. The symbols correspond to data obtained from the literature (black), the 1-m USNO telescope at Flagstaff (blue), the OAN Mexico (light blue), and FORS1/VLT (red). Lines indicate phases of power-law decay, with the first one from early data²¹ (not shown). Yellow bars mark re-brightening transitions. Contributions from an underlying supernova (solid curved line) do not become significant until ~ 10 days after the γ -ray burst.

with the model predictions^{5–7}, the position-angle changes are not, and thus it remains to be proven whether or not these early polarization data support the break at 0.4 days as being due to a jet. A connection to theoretical models throughout the entire period covered in Fig. 2 is a daunting task. A detailed comparison of our data with characteristic features of various theoretical models is beyond the scope of this paper, and will be presented elsewhere.

Thus our data, to our knowledge, constitute the most complete and dense sampling of the polarization behaviour of a γ -ray burst afterglow to date. For GRB030329 we conclude that the afterglow polarization probably did not rise above $\sim 2.5\%$ at any time, and that the polarization did not correlate with the flux. The low level of polarization implies that the components of the magnetic field parallel and perpendicular to the shock do not differ by more than $\sim 10\%$, and suggests an entangled magnetic field, probably amplified by turbulence behind shocks, rather than a pre-existing field. This is in contrast with the high level of polarization detected in the prompt γ -rays from GRB021206¹⁸ and suggests a different structure and origin of the magnetic field in the prompt versus afterglow emission regions. Evolving polarization properties provide a unique diagnostic tool for γ -ray burst studies, and the extremely complex light curve of the optical afterglow of GRB030329 emphasizes that measurements should be carried out with high sampling frequency. □

Received 16 July; accepted 30 September 2003; doi:10.1038/nature02077.

1. Stanek, K. Z. *et al.* Spectroscopic discovery of the supernova 2003dh associated with GRB 030329. *Astrophys. J.* **591**, L17–L20 (2003).
2. Hjorth, J. *et al.* A very energetic supernova associated with the γ -ray burst of 29 March 2003. *Nature* **423**, 847–850 (2003).
3. Woosley, S. E., Eastman, R. G. & Schmidt, B. P. Gamma-ray bursts and type Ic supernova SN 1998bw. *Astrophys. J.* **516**, 788–796 (1999).
4. Mészáros, P. Theories of gamma-ray bursts. *Annu. Rev. Astron. Astrophys.* **40**, 137–169 (2002).
5. Sari, R. Linear polarization and proper motion in the afterglow of beamed gamma-ray bursts. *Astrophys. J.* **524**, L43–L46 (1999).
6. Gruzinov, A. Strongly polarized optical afterglows of gamma-ray bursts. *Astrophys. J.* **525**, L29–L31 (1999).
7. Ghisellini, G. & Lazzati, D. Polarization light curves and position angle variation of beamed gamma-ray bursts. *Mon. Not. R. Astron. Soc.* **309**, L7–L11 (1999).
8. Hjorth, J. *et al.* Polarimetric constraints on the optical afterglow emission from GRB 990123. *Science* **283**, 2073–2075 (1999).
9. Wijers, R. A. M. J. *et al.* Detection of polarization in the afterglow of GRB 990510 with the ESO Very Large Telescope. *Astrophys. J.* **523**, L33–L36 (1999).
10. Covino, S. *et al.* GRB 990510: linearly polarized radiation from a fireball. *Astron. Astrophys.* **348**, L1–L4 (1999).
11. Rol, E. *et al.* GRB 990712: First indication of polarization variability in a gamma-ray burst afterglow. *Astrophys. J.* **544**, 707–711 (2000).
12. Björnsson, G., Hjorth, J., Pedersen, K. & Fynbo, J. U. The afterglow of GRB 010222: A case of continuous energy injection. *Astrophys. J.* **579**, L59–L62 (2002).
13. Masetti, N. *et al.* Optical and near-infrared observations of the GRB 020405 afterglow. *Astron. Astrophys.* **404**, 465–481 (2003).
14. Covino, S. *et al.* Polarization evolution of the GRB 020405 afterglow. *Astron. Astrophys.* **400**, L9–L12 (2003).
15. Bersier, D. *et al.* The strongly polarized afterglow of GRB 020405. *Astrophys. J.* **583**, L63–L66 (2003).
16. Barth, A. *et al.* Optical spectropolarimetry of the GRB 020813. *Astrophys. J.* **584**, L47–L51 (2003).
17. Rol, E. *et al.* Variable polarization in the optical afterglow of GRB 021004. *Astron. Astrophys.* **405**, L23–L27 (2003).
18. Coburn, W. & Boggs, S. E. Polarization of the prompt γ -ray emission from the γ -ray burst of 6 December 2002. *Nature* **423**, 415–417 (2003).
19. Vanderspeke, R., Crew, G., Doty, J., Villaseñor, J. & Monelly, G. GRB 030329 (=H2652): A long, extremely bright GRB localized by HETE WXM and SCX. *GCN Circ.* 1997 (2003).
20. Price, P. A. *et al.* The bright optical afterglow of the nearby γ -ray burst of 29 March 2003. *Nature* **423**, 844–847 (2003).
21. Uemura, M. *et al.* Structure in the early afterglow light curve of the γ -ray burst of 29 March 2003. *Nature* **423**, 843–844 (2003).
22. Greiner, J. *et al.* Redshift of GRB 030329. *GCN Circ.* 2020 (2003).
23. Kouveliotou, C. *et al.* Identification of two classes of gamma-ray bursts. *Astrophys. J.* **413**, L101–L104 (1993).
24. Burenin, R. *et al.* First hours of the GRB 030329 optical afterglow. *Astron. Lett.* **29**(9), 1–6 (2003).
25. Marshall, F. E., Markwardt, C. & Swank, J. H. GRB 030329: fading X-ray afterglow with RXTE. *GCN Circ.* 2052 (2003).
26. Tiengo, A. *et al.* The X-ray afterglow of GRB 030329. *Astron. Astrophys.* **409**, 983–987 (2003).
27. Wang, L., Howell, D. A., Höflich, P. & Wheeler, J. C. Bipolar supernova explosions. *Astrophys. J.* **550**, 1030–1035 (2001).
28. Leonard, D. C., Filippenko, A. V., Chornock, R. & Foley, R. J. Photospheric-phase spectropolarimetry and nebular-phase spectroscopy of the peculiar type Ic supernova 2002ap. *Publ. Astron. Soc. Pacif.* **114**, 1333–1348 (2002).

Acknowledgements This work is primarily based on observations collected at ESO, Chile, with additional data obtained at the German-Spanish Astronomical Centre Calar Alto, operated by the Max Planck Institute for Astronomy, Heidelberg, jointly with the Spanish National Commission for Astronomy, the NOT on La Palma, Canary Islands, and the Observatorio Astronómico Nacional, San Pedro, Mexico. We are grateful to the staff at the Paranal, Calar Alto and NOT observatories, in particular A. Aguirre, M. Alises, S. Hubrig, A. O. Jaunsen, C. Ledoux, S. Pedraz, T. Szeifert, L. Vanzi and P. Vreeswijk for obtaining the service mode data reported here.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.G. (jcg@mpe.mpg.de).

Experimental observation of symmetry-breaking nonlinear modes in an active ring

Sergej O. Demokritov¹, Alexander A. Serga¹, Vladislav E. Demidov¹, Burkard Hillebrands¹, Michail P. Kostylev² & Boris A. Kalinikos²

¹Fachbereich Physik und Forschungsschwerpunkt MINAS, Technische Universität Kaiserslautern, Erwin-Schrödinger-Straße 56, 67663 Kaiserslautern, Germany

²St Petersburg Electrotechnical University, 197376 St Petersburg, Russia

Solitons are large-amplitude, spatially confined wave packets in nonlinear media. They occur in a wide range of physical systems, such as water surfaces, optical fibres, plasmas, Bose–Einstein condensates and magnetically ordered media^{1,2}. A distinguishing feature of soliton behaviour that is common to all systems, is that they propagate without a change in shape owing to the stabilizing effect of the particular nonlinearity involved^{1,3}. When the propagation path is closed, modes consisting of one or several solitons may rotate around the ring, the topology of which imposes additional constraints on their allowed frequencies and phases^{4,5}. Here we measure the mode spectrum of spin-wave solitons in a nonlinear active ring constructed from a magnetic ferrite film. Several unusual symmetry-breaking soliton-like modes are found, such as ‘Möbius’ solitons, which break the fundamental symmetry of 2π -periodicity in the phase change acquired per loop: a Möbius soliton needs to travel twice around the ring to meet the initial phase condition.

Despite the numerous publications on nonlinear wave propagation^{1–3,6,7}, multi-soliton modes in confined nonlinear systems have received very little attention to date. Very recently Carr *et al.* analysed multi-soliton modes in a box with infinitely high walls and, equivalently, a ring^{4,5}, and they find that the symmetry of these modes can differ from that of linear modes of the system. The linear modes of a box are sinusoidal standing waves enumerated by the number of half-wavelengths that satisfy the box size. There is a direct connection between the mode number of a linear mode and its mirror symmetry with respect to the centre of the box: odd and even modes are symmetric and antisymmetric, respectively. It was shown⁵ for the case of an ‘attractive nonlinearity’ necessary for solitons to form that nonlinear modes can be classified by the number of solitons comprising a particular mode. However, there is no general connection between the symmetry properties of a mode and the number of the solitons. For example, two modes consisting of two solitons are reported: one mode is antisymmetric and the other is symmetric with respect to the box centre. The symmetry of the former mode corresponds to the symmetry of the linear mode with two half-wavelengths, while the latter mode has no linear counterpart and was therefore termed a ‘symmetry-breaking mode’.

As a model system, we constructed an active nonlinear ring based

on spin-wave propagation in a magnetic low-loss film with longitudinally applied field. This system exhibits the so-called attractive nonlinearity necessary for soliton formation, and thus the nonlinear eigenmode spectrum consists of sets of soliton-like wavepackets. (We use the term 'soliton-like mode' because the excitations found are not true solitons owing to the residual lossy character of the ring outside the location of amplification. The difference between soliton-like modes and true solitons has been discussed extensively, see refs 8, 9 for example.) The solitons are envelope solitons, that is, the envelope function of a wavepacket has a soliton shape.

The experimental arrangement is shown schematically in Fig. 1. We studied spin-wave packets travelling in a single-crystalline (111)-oriented yttrium iron garnet (YIG) film of 7- μm thickness, 1.5-mm width, and 40-mm length, forming a waveguide. Two short-circuited microstrip antennae (length 2.5 mm, width 0.05 mm) for the excitation and detection of spin waves are attached to the film at a distance $l = 8\text{ mm}$ apart from each other. To compensate for the damping of the waves and to maintain unidirectionality of the ring, a linear electronic amplifier connects the output and input antenna. Its gain is chosen such that the system is below the self-generation threshold for chaotic behaviour when the phase-sensitive amplifier discussed below is disabled. An external magnetic field of 1,870 Oe is applied along the propagation direction, thus allowing for the propagation of packets of 'backward volume' spin waves.

In the linear regime, that is, for low excitation amplitudes, the ring shows a series of resonant eigenmodes with frequencies ν in the interval 7.25–7.30 GHz, spaced by $\Delta\nu = 3.2\text{ MHz}$, and wavevectors $k \approx 100\text{--}150\text{ cm}^{-1}$. The exact frequencies of the modes are determined by the phase-quantization condition over the entire ring: $kl = 2\pi n$, with integer n . An estimation shows that the phase shift accumulated in the electronic components of the ring is negligibly

small. The time for a full circuit around the ring is $T_0 = 310\text{ ns}$.

It is necessary to have a selection mechanism to address individually each eigenmode of a nonlinear ring. One widely employed method is the use of a critical threshold mechanism: only the mode with the largest amplification gain is generated, at the expense of all others. This is commonly employed in, for example, a ring laser system. Other reported methods are the use of frequency filters or time gate filters (active mode locking)^{10–13}. As a very versatile approach for mode selection, we propose and demonstrate the use of a phase-selective amplifier. This amplifier only allows modes to exist on the active ring, which have a well-defined preset phase at the position of the amplifier in the ring. In conjunction with a simple time-gating filter and phase-sensitive mode detection, this approach provides an elegant way to study the nonlinear eigenmodes in the active ring.

To achieve phase-selective amplification we use the technique of parametric parallel pumping. A sufficiently strong microwave field with frequency 2ν can parametrically amplify a spin-wave packet of frequency ν (refs 14, 15). The pumping microwave field is sent to a microstrip resonator attached to the film centred between the two antennae as shown in Fig. 1. The width of the resonator, which determines the area of pumping, is $\delta = 0.05\text{ mm}$. Since $k\delta \ll 2\pi$, non-adiabatic pumping is achieved and results in a phase-sensitive amplification coefficient—waves having phases $\pi/4 + n\pi$ with respect to the pumping field are amplified most efficiently (see ref. 16 for details).

To provide a useful definition of the phase of soliton-like wavepackets, we refer to the case of a one-dimensional spin wave packet propagating in an unconfined magnetic medium. The dynamic magnetization $m(x, t)$ describing the packet can be written as $m(x, t) = f(x - v_g t) \exp(i\varphi(x, t))$ with $\varphi(x, t) = 2\pi\nu t - kx - \varphi_a(x)$, where f is the envelope function, v_g is the group velocity, ν and k are the carrier frequency and wavevector, respectively, and $\varphi_a(x)$ is an additional phase chirp due to nonlinear effects. It is known^{1,17} that for a soliton, $\varphi_a(x)$ is almost constant across the soliton peak. Therefore, one can characterize a propagating soliton by its phase as a whole, for example, at the maximum point of the envelope function: $\varphi(t) = \varphi(x, t)$ for $x = v_g t$. This phase can be detected by mixing the soliton pulse with a continuous-wave (c.w.) reference signal, as is shown in Fig. 1. In this case we note that as the group velocity of the waves under consideration differs considerably from the phase velocity $v_{ph} = 2\pi\nu/k$, the defined phase φ is time-dependent, that is, it changes constantly while the soliton is propagating.

Without pumping in the parametric amplifier, no nonlinear modes are observed. Pumping is applied as a sequence of microwave pulses with a duration $\tau = 15\text{--}35\text{ ns}$ and period T_p . It was found that stable excitations can be observed if T_p is commensurate with T_0 ; a deviation of only a few nanoseconds destroys the mode. For brevity, the following representative results are presented and discussed for $T_p = T_0, T_0/2$, and $T_0/3$.

We begin with the simplest case of one wavepacket in the loop, that is $T_p = T_0$. Figure 2a illustrates the waveforms of the detected eigenmodes, with curve a showing the waveform of the pumping field as a reference. By sweeping the frequency (or, equivalently, the applied magnetic field) in a narrow interval of 20 MHz (8 Oe), several nonlinear eigenmodes are subsequently observed. Without phase analysis all modes detected at different frequencies are indistinguishable and consist of a series of soliton-like pulses spaced in time by T_0 (curve b). The phase detection technique reveals two possible types of the one-soliton modes.

First, we find solitons all having the same phase, as displayed in curve c. We label them N for 'normal' solitons. Second, we find a second set of modes with the phase difference between two consecutive solitons of π , as displayed in curve d. These modes need to travel twice about the ring to meet the initial condition. Because of their similarity to the Möbius strip, we label them M for

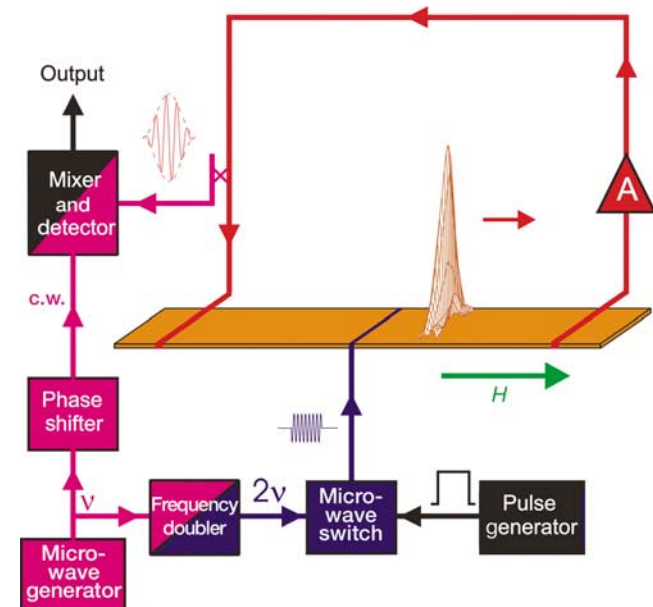


Figure 1 Schematic layout of the active ring based on an yttrium iron garnet film as the nonlinear medium for spin-wave propagation and the experimental arrangement for observation of nonlinear modes in the ring. The ring consists of the film, two antennae, and the linear microwave amplifier (A). The external magnetic field H is applied parallel to the film. Nonlinear modes of the ring are filtered using parametric pumping at double frequency at the position of the centre antenna as described in the text. The ring and the microwave components working at frequency ν are shown in red. The components working at frequency 2ν are shown in blue. The directional coupler, the phase-shifter and the mixer with detector are used for phase sensitive detection of the modes by mixing them with a c.w. reference signal.

'Möbius' mode. The waveform of the M-mode can be reversed by changing the c.w. reference phase by π (curve e). Both modes are observed alternatively at different carrier frequencies ν_N and ν_M with $\nu_N - \nu_M = 1.6$ MHz, that is, $\Delta\nu/2$.

The physical nature of the N-modes is easily understood: a soliton is propagating about the ring with the closed-loop phase shift of $2\pi n$. This set of modes, each mode characterized by n , is the nonlinear counterpart of linear excitations. The M-mode, however, has no analogue in the linear case: after completing a single loop, the phase accumulation of the soliton is $2\pi(n + 1/2)$, and the soliton needs to travel a second loop to achieve the initial state. We point out that the possible existence of Möbius modes, that is eigenmodes with a phase shift per loop that differs from an integer multiple of 2π , was not considered by Carr *et al.*^{4,5}

We also find eigenmodes consisting of several soliton-like wavepackets per loop. This happens when the pumping period is smaller than T_0 . Figure 2b displays the case of $T_p = T_0/2$, that is, a system with two equally spaced wavepackets per loop. In a manner similar to Fig. 2a, the top curve (curve a) demonstrates the pumping waveform, while the other curves are the measured waveforms of the different modes detected by phase analysis. The numbers adjacent to each curve indicate the corresponding values of the mode frequencies in units of GHz. The two vertical dashed lines indicate the time interval T_0 that each of the soliton packets needs to complete the loop.

As for the one-soliton case discussed above, these modes can be divided into normal and Möbius modes. We observe two normal modes (curves b and c) and one Möbius mode (curve d). Although the two normal modes are not completely degenerate, their frequency separation is much smaller than $\Delta\nu$. As found for the one-soliton case, the difference between the Möbius mode and both the normal modes is $\nu_N - \nu_M \cong 1.6$ MHz = $\Delta\nu/2$. By sweeping the frequency, alternating normal and Möbius modes appear, as demonstrated by curve e, which (with respect to the observed

signal) is identical to curve b, but separated in frequency by the amount $\Delta\nu$.

As a further illustration, in Fig. 2c, we demonstrate the case of three equally spaced soliton-like pulses in the ring with a temporal separation, that is, pumping period of $T_0/3$. Two normal modes and two Möbius modes are found. Again, normal and Möbius modes are observed to alternate with the same frequency interval.

Thus the phase sensitivity of the parametric pumping process constitutes a phase selection rule for the allowed eigenmodes of the ring. Owing to localized parallel pumping with frequency 2ν , the pumping gain is at maximum for phases πn , allowing for modes with a phase shift per loop of $2\pi n$ and $2\pi(n + 1/2)$. The former values correspond to normal modes and the latter to Möbius modes.

To classify the multi-soliton modes found, we quote the soliton phase relative to the phase of an arbitrary chosen soliton, which we set to zero without loss of generality. A soliton having the same phase will be indicated by '0', and one with a phase difference of π by ' π '. For $T_p = T_0$ only two one-soliton modes exist, a normal mode, (0)N, and a Möbius mode, (0)M. For the case of two-soliton modes we find, using combinatorics, four possible combinations: (00)N, (0 π)N, (00)M, and (0 π)M. However, the last two modes are physically identical, as they can be transferred into each other by shifting the origin of the timescale by $T_0/2$.

Some of the N-eigenmodes are expected to have broken symmetry, since they cannot evolve monotonously from their linear counterparts. We find that the (00)N mode has a linear counterpart, while the (0 π)N mode has the predicted broken symmetry. This is experimental verification of a symmetry-breaking eigenmode, as predicted^{4,5}. A similar analysis for $T_p = T_0/3$ predicts the existence of four non-equivalent three-soliton modes: (000)N, (0 π 0)N, (000)M, and (0 π 0)M. All these modes are observed in the experiment, as demonstrated in Fig. 2.

To good approximation, the multi-soliton eigenmodes can be

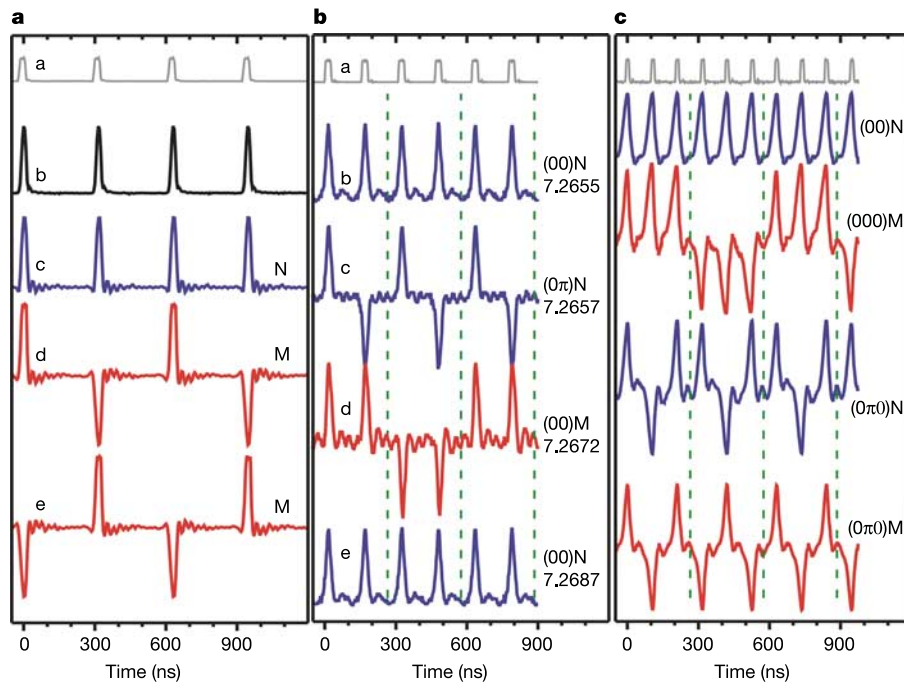


Figure 2 Waveforms of the modes observed in the ring. **a**, $T_p = T_0$, one-soliton modes. Curve a, waveform of the pumping field; curve b, waveform of a nonlinear mode obtained without phase-sensitive detection; curves c and d, waveforms of two different modes detected using phase-sensitive detection; curve e, the same as curve d with the phase of the c.w. shifted by π . **b**, $T_p = T_0/2$, two-soliton modes. Curve a, waveform of the

pumping field; curves b–e, waveforms of different modes with the numbers being the frequency of each mode in units of GHz. **c**, $T_p = T_0/3$, three-soliton modes; waveforms of the pumping field and of the observed modes. 'Normal' modes are shown in blue, and 'Möbius' modes in red. The mode nomenclature is explained in the text.

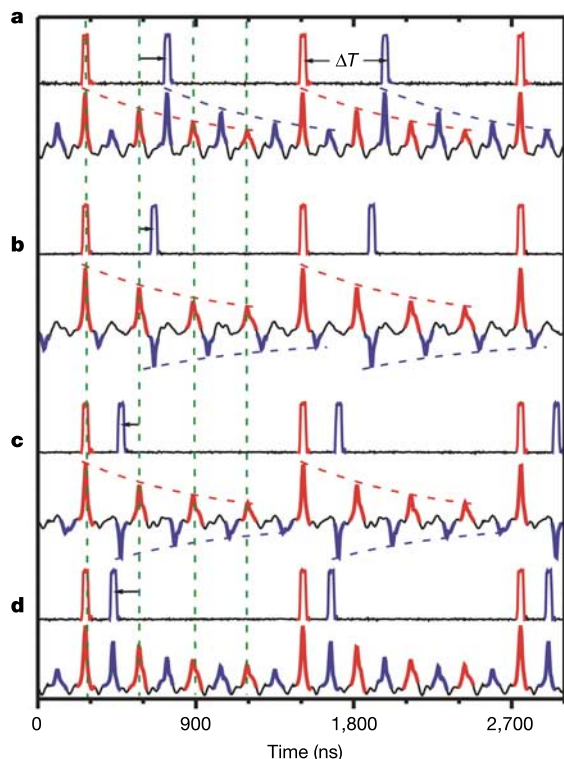


Figure 3 Waveforms of the one-soliton modes in the experiments with two sequences of the pumping pulses. Solitons corresponding to the 'red' and 'blue' pumping pulses are coloured correspondingly. The vertical green dashed lines are spaced by the time distance T_0 for reference. Red and blue dashed lines emphasize the soliton decay while it travels about the ring without amplification. Black arrows mark the time delay between the 'blue' and 'red' solitons. In all shown experiments $T_p = 4T_0$ for both sequences. Panels **a–d** correspond to the time separation between two sequences $\Delta T = 480, 400, 200$ and 150 ns, respectively. Upward waveforms represent the pumping field; downward waveforms represent the modes. For **a** and **d**, the phase difference between the two type of solitons is zero; for **b** and **c**, it is π , indicating the interaction between the solitons. Note that the time delay between the 'blue' and 'red' solitons changes with ΔT non-monotonically.

viewed as a set of soliton-like wavepackets with weak interaction in the case of a small overlap. We conclude by reporting an experiment that unambiguously demonstrates this interaction. Using appropriate time gating we prepared a mode consisting of two solitons with a variable time separation, which travel about the ring as displayed in Fig. 3. For simplicity, the pumping frequency is chosen in such a way that both solitons correspond to normal modes, that is their closed-loop phase shift is $2\pi n$. The relative phase of the first soliton (shown in red) is set to zero. Owing to phase selection imposed by the parametric pumping process, the phase of the second soliton can either be 0 or π . By reducing ΔT starting at $\Delta T = 480$ ns $\approx 3T_0/2$ (Fig. 3a) we find that the phase of the second soliton flips at $\Delta T = 400$ ns (Fig. 3b) as clear evidence of a coupling between the two solitons. The phase difference of π is still observed at $\Delta T = 200$ ns. To demonstrate that the observed phase flip is connected with the interaction between the solitons, but not with the interaction between the pumping pulses, we reduced ΔT to 150 ns (Fig. 3d). As is seen from the figure, this reduction of ΔT increases the distance between the solitons. Accordingly, the phase flip disappears. The origin of the observed interaction is not clear so far and represents a challenge for future studies.

We expect that the phenomena and concepts presented here may exist in nonlinear confined media in general. The observation of modes with broken symmetry should also be possible, for example, in optical systems. □

Received 25 April; accepted 8 September 2003; doi:10.1038/nature02042.

1. Remoissenet, M. *Waves Called Solitons: Concepts and Experiments* (Springer, Berlin/Heidelberg, 1994).
2. Khaykovich, L. *et al.* Formation of a matter-wave bright soliton. *Science* **296**, 1290–1293 (2002).
3. Agrawal, G. P. *Nonlinear Fiber Optics* (Academic, San Diego, 1995).
4. Carr, L. D., Clark, C. W. & Reinhardt, W. P. Stationary solutions of the one-dimensional nonlinear Schrödinger equation. I. Case of repulsive nonlinearity. *Phys. Rev. A* **62**, 063610 (2000).
5. Carr, L. D., Clark, C. W. & Reinhardt, W. P. Stationary solutions of the one-dimensional nonlinear Schrödinger equation. II. Case of attractive nonlinearity. *Phys. Rev. A* **62**, 063611 (2000).
6. Liu, X., Qian, L. J. & Wise, F. W. Generation of optical spatiotemporal solitons. *Phys. Rev. Lett.* **82**, 4631–4634 (1999).
7. Bauer, M. *et al.* Observation of spatiotemporal self-focusing of spin waves in magnetic films. *Phys. Rev. Lett.* **81**, 3769–3772 (1998).
8. Kivshar, Y. S. & Malomed, B. A. Dynamics of solitons in nearly integrable systems. *Rev. Mod. Phys.* **61**, 763–915 (1989).
9. Li, Z., Li, L., Tian, H., Zhou, G. & Spatschek, K. H. Chirped femtosecond solitonlike laser pulse form with self-frequency shift. *Phys. Rev. Lett.* **89**, 263901 (2002).
10. Kalinikos, B. A., Scott, M. M. & Patton, C. E. Self-generation of fundamental dark solitons in magnetic films. *Phys. Rev. Lett.* **84**, 4697–4700 (2000).
11. Kalinikos, B. A., Kovshikov, N. G., Kostylev, M. P. & Benner, H. Self-generation of spin-wave envelope soliton trains with different periods. *JETP Lett.* **76**, 253–257 (2002).
12. Kalinikos, B. A., Kovshikov, N. G. & Patton, C. E. Self-generation of microwave magnetic envelope soliton trains in yttrium iron garnet thin films. *Phys. Rev. Lett.* **80**, 4301–4304 (1998).
13. Steinmeyer, G., Sutter, D. H., Gallmann, L., Matuschek, N. & Keller, U. Frontiers in ultrashort pulse generation: Pushing the limits in linear and nonlinear optics. *Science* **286**, 1507–1512 (1999).
14. Bagada, A. V., Melkov, G. A., Serga, A. A. & Slavin, A. N. Parametric interaction of dipolar spin wave solitons with localized electromagnetic pumping. *Phys. Rev. Lett.* **79**, 2137–2140 (1997).
15. Kolodin, P. A. *et al.* Amplification of microwave magnetic envelope solitons in thin yttrium iron garnet films by parallel pumping. *Phys. Rev. Lett.* **80**, 1976–1979 (1998).
16. Melkov, G. A. *et al.* Nonadiabatic interaction of a propagating wave packet with localized parametric pumping. *Phys. Rev. E* **63**, 066607 (2001).
17. Nash, J. M., Kabos, P., Staudinger, R. & Patton, C. E. Phase profiles of microwave magnetic envelope solitons. *J. Appl. Phys.* **83**, 2689–2699 (1998).

Acknowledgements Support by the Deutsche Forschungsgemeinschaft is gratefully acknowledged. We also thank J.D. Comins for a careful reading of the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.O.D. (demokrit@physik.uni-kl.de).

Non-saturating magnetoresistance in heavily disordered semiconductors

M. M. Parish¹ & P. B. Littlewood^{1,2}

¹Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK

²National High Magnetic Field Laboratory, Pulsed Field Facility, LANL, Los Alamos, New Mexico 87545, USA

The resistance of a homogeneous semiconductor increases quadratically with magnetic field at low fields and, except in very special cases, saturates at fields much larger than the inverse of the carrier mobility, a number typically of the order of 1 T (refs 1, 2). A surprising exception to this behaviour has recently been observed in doped silver chalcogenides^{3–5}, which exhibit an anomalously large, quasi-linear magnetoresistive response that extends down to low fields and survives, even at extreme fields of 55 T and beyond. Here we present a simple model of a macroscopically disordered and strongly inhomogeneous semiconductor that exhibits a similar non-saturating magnetoresistance. In addition to providing a possible explanation for the behaviour of doped silver chalcogenides, our model suggests potential routes for the construction of magnetic field sensors with a large, controllable and linear response.

Recently, an anomalously large magnetoresistance was observed in two doped silver chalcogenides, $\text{Ag}_{2+\delta}\text{Se}$ and $\text{Ag}_{2+\delta}\text{Te}$, where the resistance displayed a positive linear dependence on the magnetic field (H) over the temperature range 4.5 K to 300 K, without any

signs of saturation at fields as high as 60 T (refs 3, 5). These characteristics make the compounds ideally suited for the development of magnetoresistive devices such as magnetic field sensors⁵, but the origin of the linear magnetoresistance still remains unclear. The silver chalcogenides are narrow-gap semiconductors⁶, so conventional theories predict that the magnetoresistance should saturate at large fields, unlike what is observed. Moreover, the silver chalcogenides possess no magnetic moments, therefore the magnetoresistance cannot be spin-mediated like the colossal magnetoresistance of the manganites⁷. Polycrystalline metals may also exhibit a linear magnetoresistance⁸, but this behaviour requires the presence of open Fermi surfaces, which is not the case here. Currently, the only proposed explanation for the silver chalcogenide phenomenon is the 'quantum magnetoresistance' of Abrikosov⁹. However, doped silver chalcogenides are granular materials⁴, and a linear magnetoresistance has also been observed in metals with surface imperfections^{10,11} and in disordered indium antimonide¹². Therefore, an alternative hypothesis is that the linear magnetoresistance of the silver chalcogenides results from large spatial fluctuations in the conductivity of the material, due to the inhomogeneous distribution of silver ions. A major advantage of such a classical theory is that the qualitative behaviour of the magnetoresistance is not sensitive to microscopic processes and, thus, only depends on temperature via changes in the macroscopic conductivity, which is consistent with experiment³.

There have been extensive theoretical investigations into the conductivity of classical, inhomogeneous media (for a summary of the results, see ref. 13), but the majority of them focus on the zero-magnetic-field case. The only known situation that yields a linear magnetoresistance, $\Delta R \equiv R(H) - R(0) \propto H$, is that of an

isotropic medium with a low volume fraction c of insulating inclusions^{14,15}, but linearity only holds for exceedingly high fields $\beta c \gg 1$ (where $\beta = \mu H$, μ is the carrier mobility and $c \ll 1$), whereas the magnetoresistance of the silver chalcogenides continues to be linear down to fields as low as 10 Oe (ref. 3). Although an effective medium approximation gives the same result for a higher volume fraction of insulating inclusions, it still only applies for high fields $\beta \gg 1$ (ref. 16). Systems with continuously varying fluctuations in the conductivity have been shown to have a non-saturating magnetoresistance in the weak disorder limit^{17,18}, but here $\Delta R \propto \beta^{2/3}$ and the calculations are once again only relevant at extremely high magnetic fields $\beta \Delta \gg 1$, where the disorder width $\Delta \ll 1$. Thus, the main limitation of the current literature is that it only deals with weak disorder, so high magnetic fields are then required to obtain a solution with non-trivial behaviour.

Here, we model a strongly inhomogeneous conductor problem by discretization into a random resistor network that we analyse numerically. Standard networks constructed from two-terminal resistors are inadequate for describing systems in magnetic fields because there are not enough terminals to take account of the Hall components, so $\Delta R = 0$. The simplest discrete model of magnetoresistance is a two-dimensional square lattice constructed of four-terminal resistors, with an external magnetic field applied perpendicular to the network. The advantage of such a network is that its sharp interfaces put it in a different class to continuous media, so that large conductivity fluctuations are already built into the system even when all the resistors are identical. Note that this network can only produce a transverse magnetoresistance, not a longitudinal one, as this requires a three-dimensional network.

We will consider the resistor unit to be a homogeneous disk with four current terminals and four voltage differences between terminals (Fig. 1a). These currents i and voltages v are connected via a 4×4 matrix z :

$$v_i = z_{ij} i_j \quad (1)$$

The matrix elements z_{ij} can be determined by solving the Laplace equation for the electric potential of a homogeneous, conducting disk, using the currents as boundary conditions, and they are found to have the form:

$$z_{ij} = \frac{\rho}{2\pi\varphi\tau} (c_{ij} + d_{ij}\beta) \quad (2)$$

where ρ is the resistivity, τ is the disk thickness, and φ is the angular width of each terminal, while c_{ij} and d_{ij} are constants dependent on the geometry. A random resistor network can thus be created by varying any of the parameters present in equation (2), although in practice we only need to consider two quantities: $r = \rho/2\pi\varphi\tau$ and μ . (To ensure that the voltages v are well defined at high magnetic field, we should require that the terminals consist of ideal thin metal contacts, $\varphi \ll 1/\beta$.)

We calculate the magnetoresistance of the network by using Kirchhoff's laws to determine the network impedance matrix Z that relates the input currents to the voltages at the input terminals, that is, $V = ZI$. As depicted in Fig. 1b, we consider the input currents to be non-zero only on two parallel edges of the network, and we also ground one of these terminals to provide a point of reference for the voltages and impose current conservation. Thus, a network of $N \times M$ resistors will possess $2N - 1$ input currents and $2N - 1$ input voltages. If we now apply a constant potential difference U across the network, the effective resistance R_{NM} of an $N \times M$ network is given by:

$$R_{NM} = \frac{U}{\sum_i I_i} = \frac{U}{\sum_i Z_{ij}^{-1} V_j} \quad (3)$$

where the sum over input currents is performed along one edge. In the limit where $N \rightarrow \infty$ for a square network, R_{NN} will give us the magnetoresistance of a strongly inhomogeneous medium.

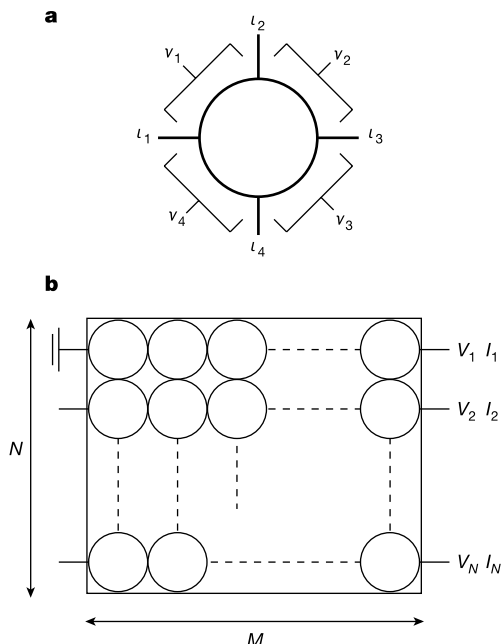


Figure 1 Four-terminal network resistor unit, and a schematic diagram of an $N \times M$ resistor network. **a**, The resistor unit is a homogeneous, conducting disk, with currents i entering the terminals and voltage differences v between the terminals. **b**, The $N \times M$ resistor network has voltage V_i and current I_i associated with the i th input terminal on one of the vertical edges, and the current is zero on the horizontal edges. One terminal is grounded to provide a reference for the voltages, and the current at this grounded terminal is ignored because of current conservation. To model the measured magnetoresistance ΔR of a macroscopic sample, a constant potential difference across the network is applied by completely grounding the left vertical edge and setting the right vertical edge to a constant voltage U .

Although equation (3) is awkward to solve analytically for large networks, some general properties of the magnetoresistance can be extracted from the symmetries of Z . It can be demonstrated that:

$$Z = S + \beta A \quad (4)$$

where S and A are symmetric and antisymmetric matrices, respectively, which are independent of β in both the limits $\beta \rightarrow 0$ and $\beta \rightarrow \infty$. This result is not surprising, because the matrix z in equation (2) also has this form when one current terminal has been grounded. Indeed, this structure can be justified on physical grounds, as an antisymmetric matrix implies dissipationless current flow, which is consistent with motion in a magnetic field. Therefore, Z becomes antisymmetric as $\beta \rightarrow \infty$. Moreover, it will be an odd antisymmetric matrix because Z is always a $(2N - 1) \times (2N - 1)$ matrix, so an eigenvalue λ_0 will approach zero at large fields. If we explicitly take out the factor β so that $Z = \beta Z_\beta$, we can write the sum of input currents in equation (3) as:

$$\sum_i I_i = \frac{1}{\beta} \sum_i \sum_n \frac{w_{n,i} w_{n,j}^T}{\lambda_n} V_j \quad (5)$$

where w_n and λ_n are the n th eigenvector and eigenvalue of Z_β , respectively. Owing to the zero eigenvalue in the denominator, the zeroth term will determine whether the current goes to zero as $\beta \rightarrow \infty$ as all other terms vanish in this limit.

For the simplest case where all resistors in the network are

identical, we obtain:

$$\sum_i I_i \propto \begin{cases} \frac{U \delta_N(\beta)}{2N-1}, & \text{if } N \text{ is even} \\ \frac{U}{2N-1}, & \text{if } N \text{ is odd} \end{cases} \quad (6)$$

where $\delta_N(\beta)$ is a high-field correction factor that vanishes as $\beta \rightarrow \infty$. Hence, $N \times M$ networks with even N will exhibit a non-saturating magnetoresistance, whereas networks with odd N will exhibit saturation except when the network is infinite.

A numerical analysis of $N \times N$ uniform networks produces magnetoresistances that all agree with equation (6). Plotted in Fig. 2a is the normalized magnetoresistance $\Delta R(H)/R(0)$ as a function of magnetic field for odd- N square networks of different sizes. The saturation level scales linearly with N as predicted by equation (6), and $\Delta R(H)/R(0) \propto \beta^2$ when $\beta < 1$, because ΔR must always be an even function of H owing to symmetry. Most importantly, we can see that the magnetoresistance curves collapse onto a straight line as $N \rightarrow \infty$, so that $\Delta R(H)/R(0) \propto \beta$ when $\beta > 1$ for infinite networks. The magnetoresistances of even- N networks also collapse onto a linear curve in the infinite size limit, as shown in Fig. 2b, but the limit is approached from above instead of below. Moreover, finite even- N networks of disks always exhibit a non-saturating magnetoresistance, unlike the single, homogeneous, van der Pauw disk with an embedded concentric inhomogeneity, which exhibits an extremely large but saturating magnetoresistance^{19,20}. This non-saturating, linear behaviour makes large networks ideal candidates for sensors of large magnetic fields.

For infinite networks, $\Delta R(H)/R(0) \propto \mu$ when $\beta > 1$, and it is independent of charge carrier density, unlike in Abrikosov's theory. This explains why $\Delta R(H)/R(0)$ of the silver chalcogenides decreases with increasing temperature, as this corresponds to a decrease in μ due to phonon excitations. In addition, it is consistent with other experiments that indicate that μ is the most important optimization variable²¹. However, the crossover from linear to quadratic behaviour occurs at $H \approx \mu^{-1}$, which is much higher than the crossover field of the silver chalcogenides.

Of course, inhomogeneous materials are generally random in nature, so it is necessary to consider random resistor networks and see whether they fundamentally change the uniform network results. In equation (4), a random resistor network corresponds to a random matrix A , and this situation is more complex than the uniform case because the zero eigenvalue is associated with a distribution of

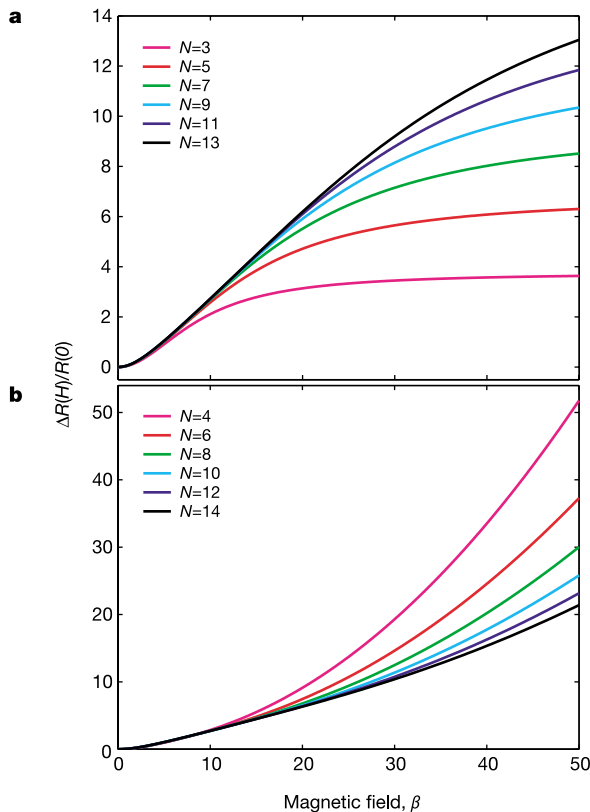


Figure 2 Normalized magnetoresistance $\Delta R(H)/R(0)$ as a function of dimensionless magnetic field β for different sized $N \times N$ uniform networks. **a**, For networks of odd N , the magnetoresistance saturates at high fields at a level that scales linearly with N , whereas at low fields, it crosses over from a linear to a quadratic dependence. As $N \rightarrow \infty$, the magnetoresistance curves collapse onto a straight line. **b**, Networks of even N always exhibit a non-saturating magnetoresistance, which collapses onto a straight line from above as $N \rightarrow \infty$.

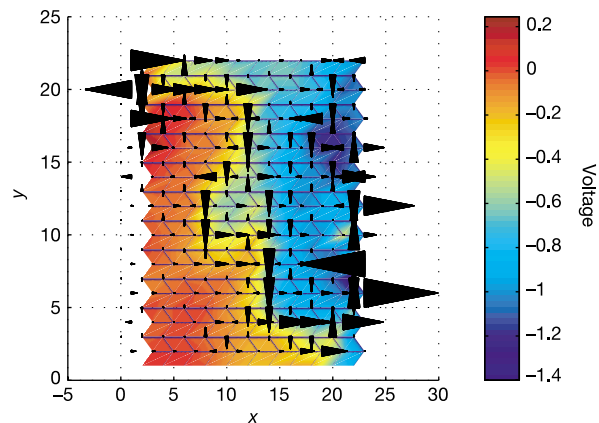


Figure 3 Visualization of currents and voltages at large magnetic field in a 10×10 random network of disks with radii 1 (arbitrary units), where the potential difference $U = -1$ V. The black arrows represent the currents, and arrow size depicts the magnitude of the current. The major current path is perpendicular to the applied voltage for a significant proportion of the time, which implies that the magnetoresistance is provided internally by the Hall effect, which is therefore linear in H .

eigenvectors when N is large. Nonetheless, we would also expect the magnetoresistance to be non-saturating as $N \rightarrow \infty$, because the distribution of eigenvector elements w_{0i} will be centred on zero, so summing over these eigenvectors in equation (5) gives zero.

Solving equation (3) numerically, we find that for finite random networks, ΔR depends on the particular network configuration, so there is a large range in behaviour, with certain configurations giving $\Delta R \propto H$ for networks as small as 4×4 . For large networks, a visualization of the current paths in the network helps us to understand the physical processes that are occurring. As shown in Fig. 3, the currents behave in a counterintuitive manner at high magnetic fields as some currents reverse direction, creating loops within the inhomogeneous system. The voltage landscape is also non-trivial, although we can see that the major current paths almost exactly follow the voltage contours, as expected at high fields. An interesting observation is that the current paths are perpendicular to the applied voltage for a significant proportion of the time. This leads us to expect a linear magnetoresistance for the network, because we would expect current flowing perpendicular to the longitudinal voltage to contribute the Hall resistance $R_h \propto H$ to the effective magnetoresistance.

As network size increases, the range in behaviour diminishes and the proportion of network configurations giving a non-saturating ΔR increases, in agreement with the analytical prediction. Furthermore, the magnetoresistance averaged over many network configurations gains a more linear dependence on magnetic field for large enough network sizes, so this, coupled with the decreasing range in behaviour, implies that $\Delta R \propto H$ for infinite random networks. Figure 4 depicts the linear, averaged $\Delta R(H)/R(0)$ of 20×20 networks for different mobility distributions, which are all assumed to be gaussian. Note that we can include positive charge carriers (holes) in this model by allowing the mobility to be positive as well as negative. At sufficiently large magnetic fields, we see that $\Delta R(H)/R(0) \propto \langle \mu \rangle$ for $\Delta \mu / \langle \mu \rangle < 1$, and $\Delta R(H)/R(0) \propto \Delta \mu$ for $\Delta \mu / \langle \mu \rangle > 1$, where $\langle \mu \rangle$ is the average mobility and $\Delta \mu$ is the width of the mobility disorder, so the magnetoresistive response is also strongly controlled by μ in the random system. In particular,

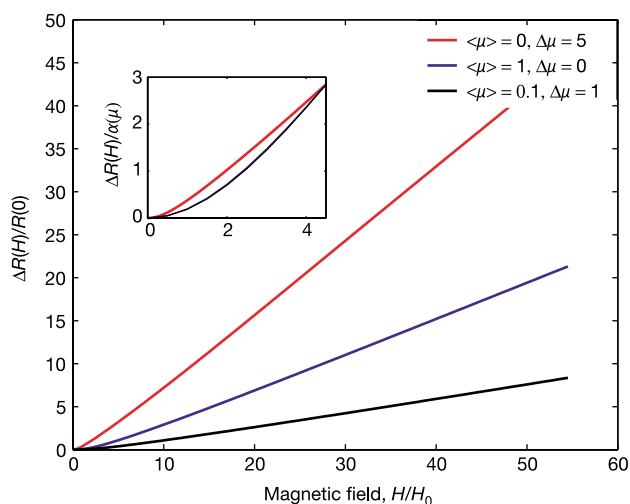


Figure 4 Average normalized magnetoresistance $\Delta R(H)/R(0)$ as a function of dimensionless magnetic field H/H_0 of 20×20 random resistor networks for different mobility distributions, where $H_0 = 1$ kOe is a typical field scale. The magnetoresistance was averaged over ten random network configurations, and the mobility distributions were taken to be gaussian and measured in units of H_0^{-1} . Strong mobility disorder results in a large magnetoresistive response. Inset, by scaling the curves so that they all have the same magnetoresistance at around 4 kOe, it can be seen that linearity continues down to lower fields when the mobility disorder is large.

the crossover field is $\langle \mu \rangle^{-1}$ for $\Delta \mu / \langle \mu \rangle < 1$ and $(\Delta \mu)^{-1}$ for $\Delta \mu / \langle \mu \rangle > 1$. Therefore, even when the characteristic field $\langle \mu \rangle^{-1}$ is of the order of 1 T, the measured crossover field of a disordered semiconductor can be several orders of magnitude smaller, provided that the mobility disorder is large. In particular, we propose that the large, linear magnetoresistance observed in silver telluride when $\langle \mu \rangle \approx 0$ (ref. 4) is due to strong disorder in the mobility.

The silver chalcogenides are promising candidates for magnetic field sensors, but the size scale of the composition fluctuations is greater than 100 nm (the bound is from small-angle neutron-scattering data; M.-L. Saboungi, C. Glinka and T. F. Rosenbaum, personal communication). This could limit their application to use as macroscopic sensors rather than nanoscopic sensors, such as read-heads, even if their sensitivity could be enhanced. However, one can, in principle, fabricate arrays of miniature disks onto any conducting material, and thus develop magnetic field sensors that are cheap to manufacture and that have responses controlled by the material mobility.

Further work is required to extend the random resistor network model to three dimensions and to obtain the exact magnetic field dependence of the magnetoresistance for the infinite random network using, for example, random matrix theory, but this model is otherwise very successful in explaining the anomalous magnetoresistance observed in silver chalcogenides. The networks display a non-saturating magnetoresistance, which is a desirable property of magnetic field sensors, they can possess a linear magnetoresistance from low to high magnetic fields like the silver chalcogenides, and the linearity at low fields increases with growing disorder. □

Received 17 July; accepted 16 September 2003; doi:10.1038/nature02073.

- Kittel, C. *Quantum Theory of Solids* (Wiley, New York, 1963).
- Smith, R. A. *Semiconductors* 2nd edn (Cambridge Univ. Press, Cambridge, 1978).
- Xu, R. *et al.* Large magnetoresistance in non-magnetic silver chalcogenides. *Nature* **390**, 57–60 (1997).
- Lee, M., Rosenbaum, T. F., Saboungi, M.-L. & Schnyders, H. S. Band-gap tuning and linear magnetoresistance in the silver chalcogenides. *Phys. Rev. Lett.* **88**, 066602 (2002).
- Husmann, A. *et al.* Megagauss sensors. *Nature* **417**, 421–424 (2002).
- Dalven, R. & Gill, R. Energy gap in β -Ag₂Se. *Phys. Rev.* **159**, 645–649 (1967).
- Tokura, Y. (ed.) *Colossal Magnetoresistive Oxides* (Gordon and Breach Science, New York, 2000).
- Kapitza, P. L. The change of electrical conductivity in strong magnetic fields. *Proc. R. Soc. Lond. A* **123**, 292–372 (1929).
- Abrikosov, A. A. Quantum magnetoresistance. *Phys. Rev. B* **58**, 2788–2794 (1998).
- Büttiker, M. Magnetoresistance of very pure simple metals. *Phys. Rev. B* **42**, 3197–3200 (1990).
- Bruls, G. J. C. L., Bass, J., van Gelder, A. P., van Kempen, H. & Wyder, P. Linear magnetoresistance caused by sample thickness variations. *Phys. Rev. Lett.* **46**, 553–555 (1981).
- Beer, A. C. *Solid State Physics Supplement 4: Galvanomagnetic Effects in Semiconductors* (Academic, New York, 1963).
- Reynolds, J. A. & Hough, J. M. Formulae for dielectric constant of mixtures. *Proc. Phys. Soc. Lond. B* **70**, 769–775 (1957).
- Balagurov, B. Ya. Conductivity of inhomogeneous media in strong magnetic fields. *Sov. Phys. Solid State* **28**, 1694–1698 (1986).
- Stroud, D. & Pan, F. P. Effect of isolated inhomogeneities on the galvanomagnetic properties of solids. *Phys. Rev. B* **13**, 1434–1438 (1976).
- Bergman, D. J. & Stroud, D. G. High-field magnetotransport in composite conductors: Effective medium approximation. *Phys. Rev. B* **62**, 6603–6613 (2000).
- Herring, C. Effect of random inhomogeneities on electrical and galvanomagnetic measurements. *J. Appl. Phys.* **31**, 1939–1953 (1960).
- Dreizin, Yu. A. & Dykhne, A. M. Anomalous conductivity of inhomogeneous media in a strong magnetic field. *Sov. Phys. JETP* **36**, 127–136 (1973).
- Solin, S. A., Thio, T., Hines, D. R. & Heremans, J. J. Enhanced room-temperature geometric magnetoresistance in inhomogeneous narrow-gap semiconductors. *Science* **289**, 1530–1532 (2000).
- Solin, S. A., Thio, T. & Hines, D. R. Controlled GMR enhancement from conducting inhomogeneities in non-magnetic semiconductors. *Physica B* **279**, 37–40 (2000).
- Ogorelec, Z., Hamzic, A. & Basletic, M. On the optimisation of the large magnetoresistance of Ag₂Se. *Europhys. Lett.* **46**, 56–61 (1999).

Acknowledgements P.B.L. thanks the NHMFL for hospitality during the final drafting of this Letter. The NHMFL is supported by the National Science Foundation, the state of Florida and the US Department of Energy. This work was supported by the Association of Commonwealth Universities and the Cambridge Commonwealth Trust.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.M.P. (mmp24@cam.ac.uk).

A polymer/semiconductor write-once read-many-times memory

Sven Möller¹, Craig Perlov², Warren Jackson², Carl Taussig² & Stephen R. Forrest¹

¹Department of Electrical Engineering and Center for Photonics and Optoelectronic Materials, Princeton University, Princeton, New Jersey 08544, USA

²Hewlett-Packard Laboratories, Palo Alto, California 94304, USA

Organic devices promise to revolutionize the extent of, and access to, electronics by providing extremely inexpensive, lightweight and capable ubiquitous components that are printed onto plastic, glass or metal foils^{1–3}. One key component of an electronic circuit that has thus far received surprisingly little attention is an organic electronic memory. Here we report an architecture for a write-once read-many-times (WORM) memory, based on the hybrid integration of an electrochromic polymer with a thin-film silicon diode deposited onto a flexible metal foil substrate. WORM memories are desirable for ultralow-cost permanent storage of digital images, eliminating the need for slow, bulky and expensive mechanical drives used in conventional magnetic and optical memories. Our results indicate that the hybrid organic/inorganic memory device is a reliable means for achieving rapid, large-scale archival data storage. The WORM memory pixel exploits a mechanism of current-controlled, thermally activated un-doping of a two-component electrochromic conducting polymer.

Ultralow-cost WORM memories have widespread uses in varied applications, such as the rapid, archival storage of video images (that is, they are similar to 'flash' memories except that the data storage is permanent)^{4–7}, where the vulnerability to breakage and the relatively high cost associated with slow and power-hungry magnetic or optical disk drives are not acceptable. In a two-dimensional WORM memory array, the memory 'pixels' are read row-by-row. This architecture requires a nonlinear electronic element, such as a p–n junction diode, at each intersection of the row and column electrodes (Fig. 1, top). Furthermore, to

distinguish between a 'written' pixel (for example, one that contains a logical '1') from an unwritten pixel (logical '0'), it is necessary to integrate a writable fuse in series with the diode. The fuse can be normally closed, or normally open (that is, an 'anti-fuse'), with the former device more suitable for testing and quality assurance of the unwritten memory. The fuse must be capable of reliable and rapid opening on the application of current or voltage of a significantly different magnitude than that used to read the memory, and it must have long-term operational stability in both the written and the unwritten states. Here we demonstrate a fused memory pixel consisting of the conductive electrochromic polymer (Baytron P, from Bayer Corp.)^{8,9}, polyethylenedioxythiophene (PEDT):poly-styrene sulphonic acid (PSS) (or 'PEDOT'), layered onto the surface of a thin-film Si p–i–n diode deposited onto a stainless steel substrate (Fig. 1, bottom).

The fused switching of the memory element (see Methods section for fabrication details) is shown in Fig. 2. The forward current of the Si diode (open circles) reaches 100 μA at a voltage of 1.5 V, whereas under reverse bias, the leakage is only 2.3 nA at -1.5 V, giving a rectifying ratio of $\sim 10^5$. Owing to the series resistance of the polymer film (of thickness $d = 40$ nm), the forward current is reduced to 400 nA for the hybrid device. Furthermore, the reverse leakage current is also increased (leading to a rectification ratio of ~ 50 , see Fig. 2), owing to the presence of a large-area, conducting polymer layered across the full Si surface. In a passive matrix memory, the Si and polymer regions between the row and column contacts are selectively removed, thus eliminating this shunt path. Indeed, removing the 50-nm-thick PEDOT layer in a fused Si device by exposure to an oxygen plasma reduced the dark current by a factor of more than 3×10^3 .

To switch the fuse state from 'closed' to 'open', the voltage is ramped at 1.0 V s^{-1} from $0 \text{ V} \rightarrow 10 \text{ V} \rightarrow 0 \text{ V}$. During each voltage step of duration $t_d = 10$ ms, the voltage is applied for $\tau = 4$ ms (a 40% duty cycle) (Fig. 3a inset). On applying this cyclic voltage ramp, the fuse opens, reducing the current to a background leakage of ≤ 10 pA at all voltages applied. Hence, the contrast between a '1' and '0' increases by $\sim 10^4$ at a forward bias of 1.5 V. In switching ~ 200 PEDOT films, the film always made the transition to the non-conducting state without a single short observed. The reliability,

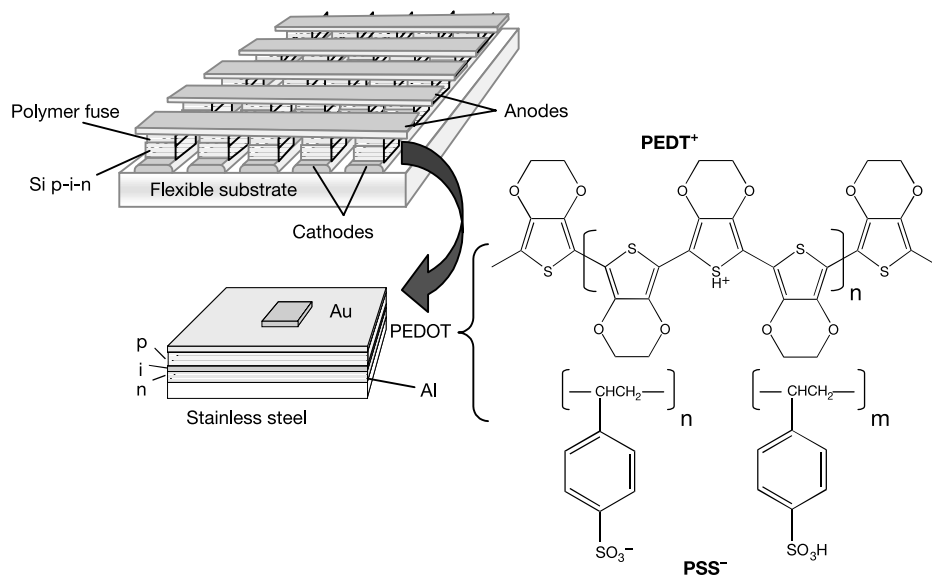


Figure 1 Generalized architecture of the WORM memory, and the materials used in its implementation. Top, conceptual view of a hybrid organic/inorganic semiconductor WORM memory. Bottom, diagram of the WORM memory element used in this study,

employing an Al cathode coated, flexible stainless steel substrate. Also shown is the chemical structural formula of the two-component conductive electrochromic polymer, PEDOT, used as the fuse material in the WORM memory element. See text for details.

irreversibility and reproducibility of the switching process are essential to a practical memory array.

To clarify the origin of the switching process, we show in Fig. 3a the current density (J) as a function of voltage (V) for a $d = 40$ nm film sandwiched between ITO (indium tin oxide) and Au contacts, where $\tau = 4$ ms, while the voltage is swept as before. The film undergoes a transition to its high-resistivity state by traversing two distinct regions: region A, from $V = 0$ to 4 V, and region B at $V \geq 4$ V. Despite the apparent complexity of the switching process, films of all thicknesses and under all bias regimes investigated qualitatively exhibit the same features in their J - V characteristics.

At the lowest voltages (0–2 V), the current density is reversible over the timescale of the experiment (minutes). At higher voltages, the current reaches a first peak, followed by a second peak, as shown in detail in Fig. 3b. The permanent conductivity changes induced by sweeps of up to 4 V are small (~ 3 –5 times lower than the unstressed film), and their magnitude depends on the pulse length, τ , as shown in Fig. 3b. The second, sharp peak appears at lower voltages, is reduced in magnitude, and broadens as τ is increased from 0.5 ms to 4 ms. However, once the film is biased into region B, a qualitatively different process is initiated, resulting in a dramatic and permanent decrease in conductivity. Indeed, once the last broad peak appears, J decreases precipitously, corresponding to a reduction in conductivity by several orders of magnitude. Hence, on the return sweep, the fuse is opened.

Figure 4a shows the transient response of a 55-nm-thick PEDOT film sandwiched between Au and ITO electrodes at several different voltages. At >5 V, there are two clearly resolved regions: a first plateau (open arrow), followed by a current peak (filled arrow). The delay between the pulse onset and the peak decreases with increasing voltage, reaching a current density of 700 A cm^{-2} after only $1 \mu\text{s}$ for a pulse height of 12 V. As in the voltage sweeps in Fig. 3, the current peak corresponds to the initiation of a permanent change in film conductivity, leading to the very-high-resistance phase of the open fuse. Switching times of $\sim 2 \mu\text{s}$ are observed in the thinnest ($d = 25$ nm) PEDOT films.

The peak delay from the voltage-pulse leading edge is linearly dependent on $1/(V - V_{\text{offset}})$, where $V_{\text{offset}} = 4.5$ V is the voltage drop across the Au contact necessary to inject an appreciable concentration of electrons. The transient response can be understood in terms of the double carrier injection previously observed in

amorphous Si p-i-n diodes¹⁰. In PEDOT, the current is primarily carried by holes¹¹. However, at $V > V_{\text{offset}}$ electrons are injected, thereby neutralizing some of the hole space charge current. This, in turn, increases the internal electric field within the polymer, leading to a further increase in hole current. Under these conditions, the hole current exceeds that in the absence of the counter electron charge by the 'double injection gain', giving rise to a current peak. At the higher pulse voltages, the peak delay decreases owing to the larger density of injected electrons. Although the current is controlled by the presence of electrons, the electrons themselves do not contribute appreciably to the current in Fig. 4a. The peak delay follows $t = d^2/\mu_n(V - V_{\text{offset}})$ as shown in Fig. 4b, leading to an electron mobility of $\mu_n = (1.7 \pm 0.4) \times 10^{-6} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, much lower than the hole mobility of $\sim 0.1 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ for this predominantly hole-transporting polymer.

Energy-barrier lowering via electron injection also suggests the origin of the permanent conductivity changes. As PSS is electrically insulating, conduction occurs by hole transport along low-resistivity percolation pathways formed by PEDT^+ chains throughout the film. The energy levels of PEDT^+ responsible for charge transport^{8,12,13} are shown in Fig. 4b, top inset. Holes can readily be injected from both Au and ITO into the highest occupied molecular orbital (HOMO) of PEDT^+ at 5.0 eV. Although the lowest unoccupied molecular orbital (LUMO) is only 0.6 eV above the HOMO, there is nevertheless a large energy barrier to overcome for electron injection. To promote injection of an electron, PEDT^+ is electrochemically reduced to PEDT^0 , and the energy level scheme changes

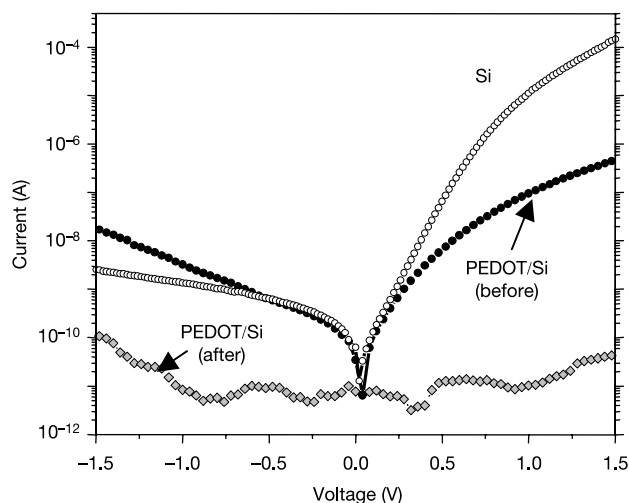


Figure 2 The switching characteristics of the WORM memory pixel. Open circles, the current/voltage characteristics of an as-deposited thin-film Si p-i-n diode; filled circles, data for an integrated 40-nm-thick PEDOT fuse/thin-film Si p-i-n diode WORM memory device in the unwritten state; diamonds, data for the WORM device in the written (blown fuse) state.

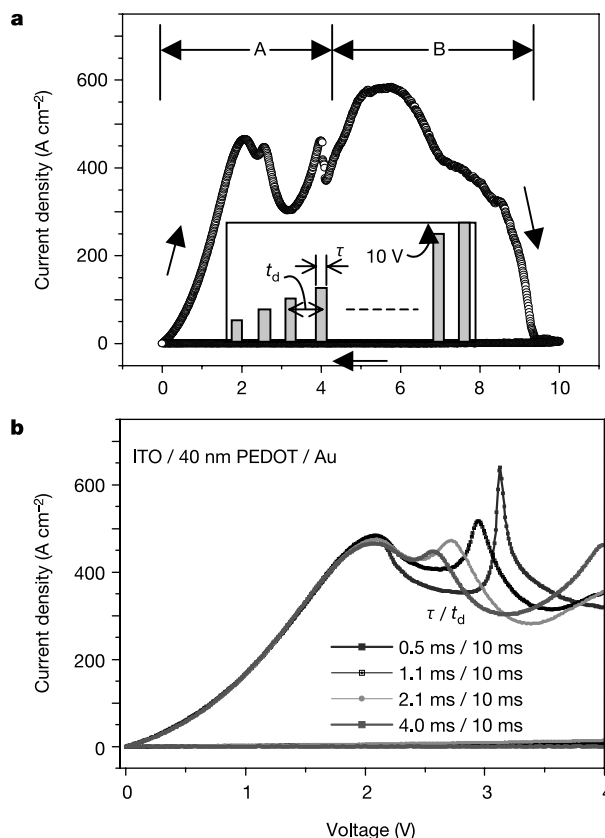


Figure 3 The switching process of the WORM memory pixel is shown under several different writing conditions. **a**, Quasi-static current density versus voltage characteristics of a 40-nm-thick PEDOT film using a pulse delay of $t_d = 10$ ms and a pulse duration of $\tau = 4$ ms. Note regions A and B. Inset, film switching sequence, defining the switching times used in the figure. **b**, Detail of the switching process occurring in region A using different τ ranging from 0.5 to 4 ms.

to that in Fig. 4b, bottom inset. In the neutral state, PEDT⁰, electrons can only occupy the LUMO at 3.9 eV, and the optical energy gap increases to 1.5 eV. Hence, there is rapid sequential oxidation and reduction by the successive transfer of holes and electrons into the polymer.

As voltage is increased beyond the first peak in region A, the sample temperature increases, leading to enhanced oxidation of PEDT⁰ accompanied by a decrease in film conductivity due to lower charge carrier concentration and/or hole mobility along the polymeric backbones. The second peak in region A shows a similar behaviour, suggesting the presence of a second species of chains with a slightly different reduction potential or chain environment within the film. Although PEDOT films are morphologically smooth, the

aqueous dispersion contains particles of different sizes consisting of interpenetrating coils of PEDT and PSS chains. The shift of the peak towards lower voltage with increasing τ is then due to slow diffusion of electrons branching into these morphologically distinct domains.

The reduction of PEDT⁺ continues with increasing current until a maximum number of chains in the neutral PEDT⁰ state are achieved for all τ , and the process saturates at voltages >2 V but still within region A. At yet higher voltages above $V_{\text{offset}} \approx 4.5$ V, electron injection leads to the process that characterizes region B—a large, permanent decrease in film conductivity by up to a factor of 10^3 . The magnitude and rapidity of the change to the low-conductivity state depends on τ and duty cycle, indicating that thermal effects contribute at high current densities. Permanent conductivity changes by thermal undoping of the polymer at elevated temperatures have been previously reported¹⁴. Calculations of the temperature rise during the current transients, based on the heat capacity and thermal conductivities typical of polymers, suggests that the maximum temperatures of $\sim 200^\circ\text{C}$ required to initiate the undoping process¹⁴ are reached at current densities of $\sim 1\text{ kA cm}^{-2}$ within the first $1\text{ }\mu\text{s}$ of the voltage pulse.

An $N \times N$ pixel matrix memory is addressed row-by-row by measuring the current flowing from N columns to the row in question. Hence, the maximum number of pixels in a row that can be accommodated without leading to read errors (that is, detecting that a pixel is written when in fact it has not been) is $N = \text{mod}[I_{\text{F1}}/I_{\text{F0}}]/S_{\text{B}} \approx 10^3$, assuming that the signal-to-background current ratio required to reliably distinguish a '1' from a '0' is $S_{\text{B}} = 10$. Here, I_{F1} and I_{F0} are the forward biased currents for the two logical states. On the basis of the performance of our devices, a 1 megabit (1 Mbit) WORM memory is therefore feasible, assuming that the current of a written pixel is $<10^{-4}$ times that of the forward-biased current of a single element in the 1 Mbit memory. Then the sum of the leakage currents from non-addressed diodes in a particular row does not significantly affect $I_{\text{F1}}/I_{\text{F0}}$; this condition is met for the devices in Fig. 2.

Such a memory block could be written within 1 s, and would occupy approximately 1 mm^2 , assuming that each pixel is 500 nm on a side, with a 50% filling of the array with memory elements. The concept is therefore adaptable to very large (up to 100 Mbit, depending on the addressing and writing schemes used) two-dimensional arrays that can be both read and written on short timescales (<1 s, using interleaved row scanning), consistent with the highest-performance archival WORM memories.

Finally, it is worth considering the advantages of the WORM over other, existing memory technologies. For example, erasable programmable read-only memory (EPROM) and flash memories rely on three-terminal crystalline Si transistor memory cells patterned using photolithography, whereas our device consists simply of a cross point with no critical alignments, deposited onto a thin film of non-crystalline Si on a foil substrate. Further, three-terminal memory cells occupy an area of $9F^2$, where F is the minimum feature size, while the cell for the WORM occupies only $4F^2$ (where $F \approx 500\text{ nm}$). Moreover, because of the higher temperatures and crystalline materials required for conventional memories, these platforms cannot be easily stacked to form a high-density three-dimensional memory—a non-existent limitation for the WORM memory. Thus, the non-crystalline organic cross point memory can exceed the potential density of these alternative, established memories in all three dimensions at lower cost. $\square$

Methods

The memory element comprises a thin-film silicon p-i-n diode and thin layer of PEDOT deposited onto the p-Si contact surface by spinning the aqueous dispersion at a rate of between 2,000 r.p.m. (resulting in a film thickness of $d = 60\text{ nm}$) and 6,000 r.p.m. (yielding $d = 25\text{ nm}$). To remove residual water, the films were baked at 140°C for 1 h in a nitrogen-filled glove box ($<1\text{ p.p.m. O}_2$, $<1\text{ p.p.m. H}_2\text{O}$). High-conductivity Baytron P Al4071 was used, with a solid content of $\sim 1.5\%$ and a PEDT:PSS weight ratio of 1:2.5. In the p-doped

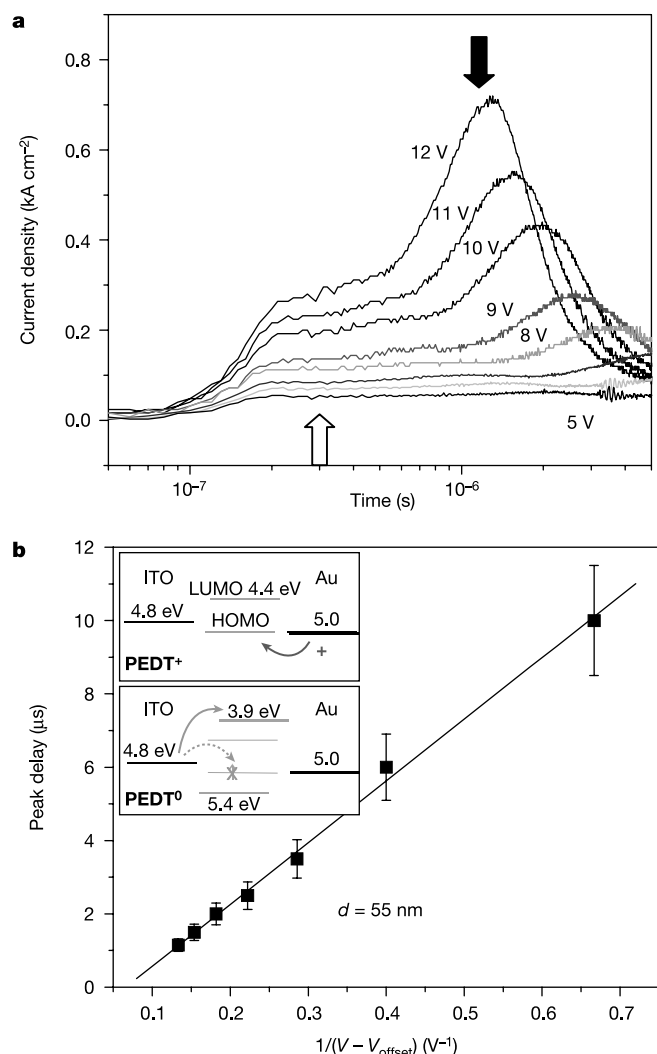


Figure 4 The behaviour of the WORM memory element under transient voltage pulse conditions. **a**, Transient response of the current density across a 60-nm-thick PEDOT film as a function of applied voltage during the pulse. The pulse duration is $>10\text{ }\mu\text{s}$, obtained using a voltage pulse generator with a rise time of 100 ns, limiting the current transient response observed at the onset of the pulse. Open arrow shows the plateau region where no changes in conductivity are observed; the filled arrow indicates the current peak corresponding to the process where there is a significant drop in conductivity, as is apparent from the slow drop in current density following the peak. **b**, Peak transient delay as a function of $1/(V - V_{\text{offset}})$, where $V_{\text{offset}} = 4.5\text{ V}$ is required for electron injection across the Au/PEDOT barrier. The straight-line fit to the data yields an electron mobility of $\mu_{\text{n}} = (1.7 \pm 0.4) \times 10^{-6}\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$. Inset, the energy level diagrams of PEDT in the oxidized (upper panel) and neutral (lower panel) states, along with the electrode arrangement in the films used in Figs 3 and 4.

state, PEDOT films exhibit conductivities of up to 1 S cm^{-1} , with a principal absorption edge at 0.6 eV, making it optically absorbing in the infrared but transparent across the visible spectrum^{8,9}. The thin-film Si diodes were fabricated by low-temperature deposition on a lightweight, stainless steel substrate coated with a 100-nm-thick Al layer that served as the cathode. The p-i-n rectifying structures consisted of a grown, 15-nm $n^+ \text{Si}/400\text{-nm}$ undoped (intrinsic) i-Si/25-nm $p^+ \text{Si}$ junction region. To independently investigate the switching properties of the polymer, an ITO/polymer/metal sandwich device was fabricated on precleaned, 20 Ω per square, ITO-coated glass substrates. All devices are contacted on the polymer surface via thermally evaporated, $17 \mu\text{m}^2$ Au electrodes. The devices were electrically characterized in air. In our temperature simulations, the thermal impedance between the film and substrate is neglected. This implies that within the first ~ 100 ns of the voltage onset, the substrate/film interface remains at room temperature.

Received 15 April; accepted 12 September 2003; doi:10.1038/nature02070.

- Forrest, S. R., Burrows, P. E. & Thompson, M. E. Organic emitters promise a new generation of displays. *Laser Focus World*, Feb., 99–101 (1995).
- Gelinck, G. H., Geuns, T. C. T. & Leeuw, D. M. d. High-performance all-polymer integrated circuits. *Appl. Phys. Lett.* **77**, 1487–1489 (2000).
- Peumans, P., Yakimov, A. & Forrest, S. R. Small molecular weight organic thin-film photodetectors and solar cells. *J. Appl. Phys.* **93**, 3693–3723 (2003).
- Saluel, D., Daval, J., Bechevet, B., Germain, C. & Valon, B. Ultra high density data storage on phase change materials with electrical micro-tips. *J. Magn. Magn. Mater.* **193**, 488–491 (1999).
- Asokan, S. Electrical switching in chalcogenide glasses—Some newer insights. *J. Optoelectron. Adv. Mater.* **3**, 753–756 (2001).
- Stocker, H. J. Bulk and thin film switching and memory effects in semiconducting chalcogenide glasses. *Appl. Phys. Lett.* **15**, 55–57 (1969).
- Hua, Z. Y., Chen, G. R., Xu, W. & Chen, D. Y. New organic bistable films for ultrafast electric memories. *Appl. Surf. Sci.* **169**, 447–451 (2001).
- Heuer, H. W., Wehrmann, R. & Kirchmeyer, S. Electrochromic window based on conducting poly(3,4-ethylenedioxythiophene)poly(styrene sulfonate). *Adv. Funct. Mater.* **12**, 89–94 (2002).
- Groenendaal, B. L., Jonas, F., Freitag, D., Pielartzik, H. & Reynolds, J. R. Poly(3,4-ethylenedioxythiophene) and its derivatives: Past, present, and future. *Adv. Mater.* **12**, 481–494 (2000).
- Hack, M. & Street, R. A. Analysis of double injection in amorphous silicon p-i-n diodes. *J. Appl. Phys.* **72**, 2331–2339 (1992).
- Johansson, T., Pettersson, L. A. A. & Inganäs, O. Conductivity of de-doped poly(3,4-ethylenedioxythiophene). *Synth. Met.* **129**, 269–274 (2002).
- Greczynski, G. et al. Photoelectron spectroscopy of thin films of PEDOT-PSS conjugated polymer blend: a mini-review and some new results. *J. Electron Spectrosc. Rel. Phenom.* **121**, 1–17 (2001).
- Xing, K. Z., Fahlman, M., Chen, X. W., Inganäs, O. & Salaneck, W. R. The electronic structure of poly(3,4-ethylene-dioxythiophene): Studied by XPS and UPS. *Synth. Met.* **89**, 161–165 (1997).
- Pei, Q. B., Zuccarello, G., Ahlskog, M. & Inganäs, O. Electrochromic and highly stable poly(3,4-ethylenedioxythiophene) switches between opaque blue-black and transparent sky blue. *Polymer* **35**, 1347–1351 (1994).

Acknowledgements We thank A. Elschner and H. C. Starck for samples of Baytron P, and M. Thompson for discussions. We also thank National Renewable Energy Laboratories for supplying the thin-film Si diodes. This work was supported by HP and the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.R.F. (forrest@princeton.edu).

Rapid body size decline in Alaskan Pleistocene horses before extinction

R. Dale Guthrie

Institute of Arctic Biology, University of Alaska, Fairbanks, Alaska 99775, USA

About 70% of North American large mammal species were lost at the end of the Pleistocene epoch¹. The causes of this extinction—the role of humans versus that of climate—have been the focus of much controversy^{1–6}. Horses have figured centrally in that debate, because equid species dominated North American late Pleistocene faunas in terms of abundance, geographical distribution, and species variety, yet none survived into the Holocene epoch. The timing of these equid regional extinctions and accompanying evolutionary changes are poorly known. In an attempt to document better the decline and demise of two Alaskan Pleistocene equids, I selected a large number of fossils from the latest

Pleistocene for radiocarbon dating. Here I show that horses underwent a rapid decline in body size before extinction, and I propose that the size decline and subsequent regional extinction at 12,500 radiocarbon years before present are best attributed to a coincident climatic/vegetational shift. The present data do not support human overkill¹ and several other proposed extinction causes^{2,3}, and also show that large mammal species responded somewhat individually to climate changes^{4–6} at the end of the Pleistocene.

Cabaloid horses in Alaska were probably part of the diverse Eurasian species complex of the stocky *Equus ferus*, judging from their fossil anatomy^{7,8} and preserved mitochondrial DNA⁹. In contrast, fossils of the other equid Alaskan species have unusually long gracile metapodials, making them similar to the living *Equus hemionus* from central Asia, and also similar to fossil New World hemione-like forms. So little work has been done with these northern hemione-like equid fossils that I shall call these hemionids for simplicity, but note that their specific relationships remain uncertain.

Alaskan hemionid fossils had not previously been radiocarbon dated. They are comparatively rare in museum collections. In this study, 19 such fossils were dated, and the fact that none of the hemionid metacarpals dated after 31,000 radiocarbon years before present (31 kyr BP) makes their extinction before the Last Glacial Maximum (LGM) seem likely. It is possible that the disappearance of this small-bodied hemionid species somehow affected selection pressures on later Alaskan cabaloid horses, promoting reductions in body size and increased numbers of the latter during the LGM, around 18–20 kyr BP, but at present this seems unlikely given the implied time lag between the hemionid extinction and cabaloid response.

Compared with the hemionids, cabaloid horse fossils from Alaska are much more numerous and much better studied. Previous dates on over 220 fossil specimens from Alaska and the Yukon Territory^{10,11} revealed no dates later than 12.5 kyr BP. Other studies from Eurasia and Beringia had indicated a size decline among horses at some time during the late Pleistocene^{12,13}, but the exact timing was uncertain. Unlike previous studies, I chose to date metacarpals exclusively in order to better assess any subtle changes in body size with time and how this related to extinction. The dramatic size changes recorded just before extinction were most unexpected (Fig. 1). Explanations for these apparent size changes include the possibility that my sampling methods may somehow have affected this plot and artificially created a drop in apparent metacarpal size. This seems unlikely (see Methods). I suspect that the gap in dates at 28–35 kyr BP may reflect lower population numbers during this time, because previous, non-metacarpal, radiocarbon work on Alaskan horse fossils also showed comparatively few dates during this pre-LGM period¹⁴.

A second possibility is that a new species or morphotype of smaller horses immigrated into Alaska, creating the illusion of size decline. The apparent high metacarpal variability during the LGM (Fig. 1) as opposed to post-LGM suggests such an interpretation. Some of this variability may simply be due to the greater abundance of LGM dates. In addition, species in transition from one adaptive peak to another, undergoing rapid evolution, are often more variable¹⁵. The graph of size/time does not show two overlapping clusters, as one would expect with species replacement. Instead, smaller variants continue to emerge on the graph over time, and larger variants continue to fall out. This would necessitate multiple species replacements. In contrast, the explanation of a simple body-size decline in one widespread population seems much more parsimonious. Comparable size declines at the end of the Pleistocene are not unique to horses¹⁶. Bison declined more dramatically in an even shorter time span¹⁷ but at a later date.

The significance of this size decline among Alaskan horses just before their regional extinction is that environmental pressures provoking smaller body size may well have been the same ones that

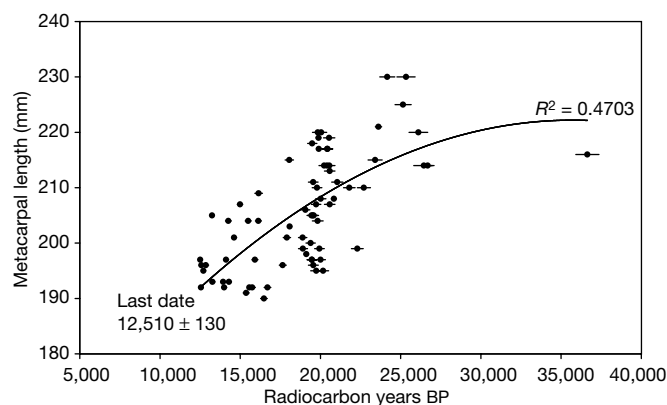


Figure 1 Length of Alaskan Pleistocene horse metacarpals plotted against AMS radiocarbon dates, with polynomial trendline. The plot suggests declining body size with time, and local extinction in Alaska around 12.5 kyr BP. Error bars are 1σ . The latest dates for horses in this study are: AMNH134BX36, Upper Cleary Creek, AA26830, $12,482 \pm 80$ (21.2); AA26819, $12,510 \pm 130$ (-21.1); AMNH6159, Fox, Alaska (AK), AMNH144-9422, Ester, AK, AA26810, $12,560 \pm 170$ (-21.6); AA26828, $12,580 \pm 140$ (-21.2); AMNH126-6846, Lower Goldstream Creek, AA2682, $12,710 \pm 170$ (-21.2); AMNH4339, Fairbanks Creek, AA26829, $12,860 \pm 140$ (-21.2). Data are as follows: AMNH (Frick Collection) are museum collection numbers, and AA numbers are Arizona AMS dating laboratory designations; then date is given in uncalibrated radiocarbon years with 1σ ; and then (in parentheses) $\delta^{13}\text{C}$ values (in per mil) are given. These and other radiocarbon data on Alaskan horses and hemionids are available as Supplementary Information.

ultimately resulted in their extinction.

Details of extinction timing are crucial in assessing competing theories about Pleistocene extinctions. Models positing explosive events, such as cross-species virulent disease, require the disappearance of species to coincide fairly closely, at least on a regional basis. Significant disjunctions in the timing of extinctions are also a challenge to theories that portray numerous species being rapidly 'overhunted' by a colonizing wave of human hunters. The apparent extinction of Alaskan hemionids (youngest date at $31,400 \pm 1,200$ yr BP), coinciding with the early phases of the last glacial, is so early that few researchers would argue that humans were involved. But even if we discount the Alaskan hemionid event as 'background extinction', we cannot do that for the caballoid horses. Horses figure too importantly in extinction models.

Could the extinction of Alaskan horses and their apparent size decline before extinction be the result of human hunting¹⁸? Note that this new array of Alaskan dates shows a gap of 1,000 radiocarbon years (close to 1,300 calendar years) between the last dated horse and mammoth fossils (Fig. 2). There is accumulating archaeological evidence of human presence in Alaska beginning around 12 kyr BP or slightly later¹⁹. So with Alaskan mammoths persisting at least until 11.5 kyr BP, it now appears that humans and mammoths coexisted for at least a few hundred years. A vastly longer overlap occurred across much of Asia²⁰ and Europe²¹.

In contrast, there is a hiatus of over 500 radiocarbon years between the last dated Alaskan horses and the earliest undisputed human artefacts. Unless unquestionable human archaeological sites are found in Alaska that date to the LGM, neither the rapid body diminution of Alaskan caballoid horses beginning at least around the LGM, nor the apparent decline in their relative numbers soon after LGM, can be reasonably attributed to the impact of selective human hunting on the megafauna, as some have suggested¹⁸. The present Alaskan data, showing that mammoths, the keystone grazer², survived more than a millennium after horses, runs counter to extinction scenarios based on ecological repercussions following human over-hunting of keystone grazer species^{2,22}.

Even if we allow that human hunters may have been present in

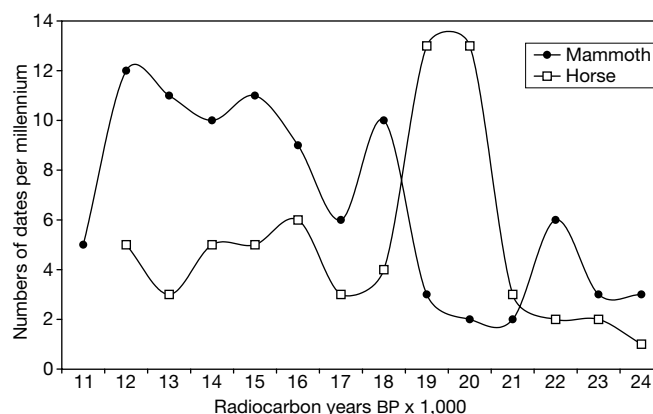


Figure 2 Smoothed plot of new Alaskan dates for horses ($n = 73$) and mammoths ($n = 72$) in radiocarbon years. Plot covers the past 24,000 years (dates summed within 1,000-year intervals). The clusters of dates, which are potentially the result of higher population densities, change significantly with time. The serial correlations of lagged values³¹ were 0.57 for horse, and 0.17 for mammoth. The latest mammoth dates in this study are: AK A-1011, Galena, AK, AA22573, $11,500 \pm 160$ (-22.9); Charles Holmes (Swan Point Site), Delta, AK, AA17601, $11,540 \pm 140$ (-20.8); NMC25941, Dawson area, Yukon Territory (YK), AA17559, $11,860 \pm 120$ (-21.2); UNM6684, Cape Lisburn, AK, AA26006, $11,910 \pm 130$ (-21.2); CMN777, YK 1915, AA17562, $11,990 \pm 130$ (-21.3). See Fig. 1 legend for data presentation. The mammoth dates are drawn from part of a more extended data set (R.D.G., manuscript in preparation).

Alaska at levels below archaeological visibility when horses became extinct at 12.5 kyr BP, the idea that such a low density of hunters could have caused horse extinction requires an unlikely scale of overkill performance. That overkill theory must further address the paradox of why super-able horse hunters should coexist with Alaskan mammoths for another 1,300 calendar years.

What, on the other hand, do we know about environmental changes in Alaska at the time of these extinctions? Broadly speaking, the last glacial was a time in which the cold/arid northern Mammoth Steppe was most extreme, though still capable of supporting a rich diversity of large mammals²². Dune fields, thick loess deposits and occasional lakes from this period point to arid and windy conditions with a treeless, short grass-sedge-sage sward^{14,23,24}. Although the region's large mammals were evidently adapted to handle cold/arid extremes, each species was evolutionarily fine-tuned to different optimal diets and habitats. Key Mammoth Steppe grazers—bison and woolly mammoth—were widespread during the LGM, though fossil evidence suggests that their relative abundance may have declined during this climatic extreme¹¹.

Proxy data show the ecological transition to Holocene conditions did not occur smoothly. A dramatic pollen shift occurred around 13–12.5 kyr BP^{25–27}; this was also a pivotal time for large mammals²⁸ and for invertebrates, like insects²⁹. Landscape changes included the creation of lakes, bogs, shrub tundra, forests, soil paludification, low-nutrient soils, and plants highly defended against herbivory. Vegetation in the north now supports a relatively small biomass of large herbivores, and almost no grazers. Northern tundra is especially marginal for significant numbers of wild horses¹⁴.

The greater abundance of both fossil bones and radiocarbon dates for horses in relation to mammoths (Fig. 2), and most other grazers, across the LGM period in Alaska is noteworthy. Virtually all the fossils in question come from the same sites, so changes in abundance between species may well be a very rough indicator of relative abundance rather than a taphonomic artefact. Horses are almost obligatory grazers, and these northern horses appear to have been the most specialized grassland-dependent component of the large mammal fauna.

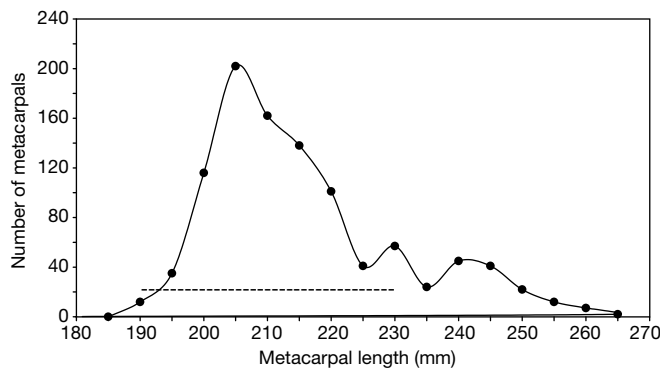


Figure 3 Smoothed plot of numbers of Alaskan horse metacarpals available in collections against their lengths summed in 5-mm increments. The dashed line represents the span from which I drew my sample (190 to 230 mm).

The present data suggest that Alaskan horses prospered during the LGM, and appear to have been particularly well adapted to the more intense versions of the cold/arid Mammoth Steppe. The data further suggest that Alaskan horses were the first steppe species to drop out at 12.5 kyr BP as the Mammoth Steppe began to unravel and be replaced by more mesic vegetation. I further hypothesize that some of the diminution observed in these horse fossils during post-LGM times can be attributed to the shift away from intensely arid steppe conditions. That woolly mammoth, a slightly less obligate steppic species, should persist in Alaska for over another millennium is congruent with this hypothesis. Perhaps the declining body size of Alaskan horses and their extinction relate not only to the absolute decline of their access to optimal food resources, but also to increasing competition with other large mammals possessing the physiological capacities to thrive on the vegetation characteristic of this 13–10 kyr BP, end-Pleistocene, transition.

As radiocarbon chronologies and proxy data of palaeoclimate gain refinement, our reconstructions of this dynamic end-Pleistocene time have become less simple. Although global climatic drivers were at play in the terminal Pleistocene, some of their effects are remarkably particular. For example, during this time we note the general replacement of caecalids—species that use a large caecum gut-diverticulum in food processing (for example, woolly mammoth, woolly rhino and horses)—by ruminants (for example, bison, moose and reindeer)^{4,10,11}. The patterns of the radiocarbon dates reported here highlight how some Alaskan large mammal species responded to changes during the Pleistocene–Holocene transition in different ways, and suggest that this individualistic response is a general phenomenon—that within a general framework, ultimately, each species had its own unique story. □

Methods

Alaskan fossils of late Pleistocene large mammals are generally collected from thawed permafrost (some bones still contain white marrow), and thus have a very high degree of collagen preservation for reliable dating. Dates were run at the US National AMS (Accelerator Mass Spectrometer) Radiocarbon Laboratory at Tucson, Arizona, with their usual pretreatment¹⁰. I chose to use metacarpals in this study for several reasons. Metacarpals are recognized as one of the best postcranial bones for distinguishing co-existing equid species, such as hemionines and caballines⁷. The cortex of these bones is so unusually dense that they are rarely used by scavengers, which probably helps make them abundant in the fossil record. Metacarpals are a critical weight-carrying structure, so their size is a good indicator of horse body mass (in this study, metacarpal width declined in the same pattern as metacarpal lengths.) Unlike woolly mammoths and most other Pleistocene large mammals, horses are barely sexually dimorphic, so I did not have to address the complex issue of sexual identity. Also unlike mammoths, bone growth among horses is quite determinate, with epiphyses fusing at adulthood, so maturity could be assessed by this fusion—only metacarpals of mature animals were used.

To make best use of funding allocated for radiocarbon dating, I did not date the largest metacarpals (Fig. 3). Earlier studies^{12,13} had shown that the largest metacarpals, ranging from 230 to 260 mm in length, tended to fall beyond radiocarbon dating range. These larger metacarpals are normally associated in Alaska with sites where Early to Middle Pleistocene sediments, as dated by tephra horizons, predominate (Cripple Creek Sump,

Gold Hill). Two of my dates of 230-mm metacarpals confirmed these previous assessments, and I deleted these specimens with infinite dates from the above plots.

The majority of Alaskan horse metacarpals in museum collections are under 230 mm. These constitute some 82% of 1,020 specimens that I examined. I selected horse metacarpals at random from this class. If small animals (<210 mm, say) had existed before the LGM they would have been captured by the sampling technique, as would have larger animals (>210 mm) that existed after the LGM. Had no actual size decline been occurring, all the means of dated cohorts before, during and after the LGM should theoretically be similar. This was not the case. The 19 Alaskan hemionine metacarpals were selected at random from the small numbers in museum collections.

Received 4 February; accepted 23 September 2003; doi:10.1038/nature02098.

- Martin, P. S. in *Quaternary Extinctions* (eds Martin, P. S. & Klein, R. G.) 345–403 (Univ. Arizona Press, Tucson, 1984).
- Zimov, S. A. *et al.* Steppe–tundra transition: A herbivore driven biome shift at the end of the Pleistocene. *Am. Nat.* **146**, 765–794 (1995).
- MacPhee, R. D. E. & Marx, P. A. in *Natural Change and Human Impact in Madagascar* (eds Goodman, S. M. & Patterson, B. D.) 169–217 (Smithsonian Institution Press, Washington, 1997).
- Guthrie, R. D. in *Megafauna and Man* (eds Agenbroad, L. D., Mead, J. I. & Nelson, L. W.) 42–54 (The Mammoth Hot Springs Scientific Papers, Vol. 1, North Dakota, 1984).
- Graham, R. W. & Lundelius, E. L. Jr in *Quaternary Extinctions* (eds Martin, P. S. & Klein, R. G.) 223–250 (Univ. Arizona Press, Tucson, 1984).
- Guthrie, R. D. in *Quaternary Extinctions* (eds Martin, P. S. & Klein, R. G.) 259–298 (Univ. Arizona Press, Tucson, 1984).
- Eisenmann, V. & Karchoud, A. Analyses multidimensionnelles de métapodes d'Equus sensu lato (Mammalia, Perissodactyla). *Bull. Mus. Natl. Hist. Nat. Paris Ser. 4*, 75–103 (1982).
- Forsten, A. *Equus lambei* the Yukon Canada wild horse not ass. *J. Mamm.* **67**, 422–423 (1982).
- Vilà, C. *et al.* Widespread origins of domestic horse lineages. *Science* **291**, 474–477 (2001).
- Harington, C. R. *Annotated Bibliography of Quaternary Vertebrates of Northern North America* (Univ. Toronto Press, Toronto, 2003).
- Matheus, P., Kunz, M. & Guthrie, R. D. in *Abstr 3rd Int. Mammoth Conf.* 3–5 (Yukon Territorial Government, Dawson, Canada, 2003).
- Forsten, A. Size decrease in Pleistocene–Holocene true or caballoid horses in Europe. *Mammalia* **55**, 407–420 (1991).
- Guthrie, R. D. in *People and Wildlife in Northern North America* (eds Gerlach, C. & Murray, M.) 32–49 (Br. Archaeol. Rep., Int. Ser. 944, 2001).
- Guthrie, R. D. *Frozen Fauna of the Mammoth Steppe* (Univ. Chicago Press, Chicago, 1990).
- Guthrie, R. D. Variability in characters undergoing rapid evolution, an analysis of *Microtus* molars. *Evolution* **19**, 214–233 (1965).
- Guthrie, R. D. in *Contributions in Vertebrate Paleontology* (eds Genoways, H. H. & Dawson, M. R.) 482–510 (Carnegie Museum of Natural History Special Publication 8, Pittsburgh, 1964).
- Wilson, M. C. Late Quaternary vertebrates and the opening of the ice-free corridor, with special reference to bison. *Quat. Int.* **32**, 97–105 (1996).
- McDonald, J. in *Quaternary Extinctions* (eds Martin, P. S. & Klein, R. G.) 409–439 (Univ. Arizona Press, Tucson, 1984).
- Holmes, C. E. Tanana River Valley archaeology circa 14,000 to 9,000 B.P. *Arctic Anthropol.* **38**, 154–170 (2001).
- Kuznetsova, T. V., Sulerzhitsky, L. D. & Siegert, Ch. in *The World of Elephants* (eds Cavarretta, G., Gioia, P., Mussi, M. & Palombo, M. R.) 289–292 (Università degli Studi de Roma “La Sapienza”, 2001).
- Stuart, A. J., Sulerzhitsky, L. D., Orlova, L. A., Kuzmin, Y. V. & Lister, A. M. The latest woolly mammoths (*Mammuthus primigenius* Blumenbach) in Europe and Asia: A review of the current evidence. *Quat. Sci. Rev.* **21**, 1559–1569 (2002).
- Aloy, J. A multispecies overkill simulation of the end-Pleistocene megafaunal mass extinction. *Science* **292**, 1893–1896 (2001).
- Guthrie, R. D. Origin and causes of the mammoth steppe: A story of cloud cover, woolly mammoth tooth pits, buckles, and inside-out Beringia. *Quat. Sci. Rev.* **20**, 549–574 (2001).
- Goetcheus, V. G. & Birks, H. H. Full Glacial upland tundra vegetation preserved under tephra in the Beringia National Park, Seward Peninsula, Alaska. *Quat. Sci. Rev.* **20**, 135–147 (2001).
- Anderson, P. M. Late Quaternary pollen records from the Kobuk and Noatak River drainages, northwestern Alaska. *Quat. Res.* **29**, 263–276 (1988).
- Lozhkin, A. V. *et al.* Late Quaternary lacustrine pollen records from Southwestern Beringia. *Quat. Res.* **39**, 314–324 (1993).
- Lamb, H. F. & Edwards, M. E. in *Vegetation History* (eds Huntley, B. & Webb, T. III) 519–556 (Kluwer Academic, Dordrecht, 1988).
- Sher, A. V. Past and future rapid environmental changes. *Glob. Environ. Change* **1** 47, 319–339 (1997).
- Lowe, J. *et al.* Climatic changes in areas adjacent to the North Atlantic during the last Glacial–Interglacial transition (14–9 ka BP): A contribution to the IGCP-235. *Quat. Sci.* **9**, 185–198 (1994).
- Brown, T. A., Nelson, D. E., Vogel, J. S. & Southon, J. R. Improved collagen extraction by Longin method. *Radiocarbon* **30**, 171–177 (1988).
- Kachigan, S. K. *Multivariate Statistical Analysis* (Radius, New York, 1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements I am grateful to R. Tedford for sampling access to these Alaskan equid and mammoth fossils. Similar thanks go to R. Gangloff and R. Harington for dating access to mammoth fossils from Alaska and the Yukon Territory. M. L. Guthrie, A. Lister and P. Matheus made suggestions that improved the manuscript. AMS dates were done at the AMS Laboratory, University of Arizona. This work was supported by the National Science Foundation.

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and request for materials should be addressed to the author (ffrdg@aurora.alaska.edu).

Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation

Jorge Tovar¹, Gloria León-Avila¹, Lidya B Sánchez^{2*}, Robert Sutak³, Jan Tachezy³, Mark van der Giezen¹, Manuel Hernández¹, Miklós Müller² & John M. Lucocq⁴

¹School of Biological Sciences, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK

²The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

³Department of Parasitology, Faculty of Science, Charles University, 12844 Prague 2, Czech Republic

⁴School of Life Sciences, WTB/MSI complex, University of Dundee, Dundee DD1 5EH, UK

* Present address: Public Health Research Institute, Newark, New Jersey 07103, USA

Giardia intestinalis (syn. *lamblia*) is one of the most widespread intestinal protozoan pathogens worldwide, causing hundreds of thousands of cases of diarrhoea each year¹. *Giardia* is a member of the diplomonads, often described as an ancient protist group whose primitive nature is suggested by the lack of typical eukaryotic organelles (for example, mitochondria, peroxisomes), the presence of a poorly developed endomembrane system and by their early branching in a number of gene phylogenies^{1,2}. The discovery of nuclear genes of putative mitochondrial ancestry in *Giardia*^{3–7} and the recent identification of mitochondrial remnant organelles in amitochondrial protists such as *Entamoeba histolytica*^{8,9} and *Trachipleistophora hominis*¹⁰ suggest that the eukaryotic amitochondrial state is not a primitive condition but is rather the result of reductive evolution. Using an *in vitro* protein reconstitution assay and specific antibodies against IscS and IscU—two mitochondrial marker proteins involved in iron-sulphur cluster biosynthesis—here we demonstrate that *Giardia*

contains mitochondrial remnant organelles (mitosomes) bounded by double membranes that function in iron-sulphur protein maturation. Our results indicate that *Giardia* is not primitively amitochondrial and that it has retained a functional organelle derived from the original mitochondrial endosymbiont.

The assembly and maturation of iron-sulphur (Fe-S) proteins is a recently identified critical function of the mitochondrion¹¹. Proteins containing Fe-S centres are widely distributed in nature and operate in metalloenzyme catalysis and electron transport. Although several Fe-S proteins are important in energy metabolism in amitochondriate organisms (for example, pyruvate:ferredoxin oxidoreductase, ferredoxin, hydrogenase)¹², almost nothing is known about the maturation of these proteins into functional enzymes or about the biosynthesis of their essential Fe-S reactive centres. Genes encoding the soluble enzyme cysteine desulphurase (IscS), a central component of the Fe-S cluster assembly system of prokaryotic and eukaryotic cells, were recently cloned from the amitochondrial protists *Trichomonas vaginalis* and *Giardia intestinalis*⁵. Because phylogenetic analyses of these genes support their mitochondrial ancestry, it is possible that their encoded proteins are still compartmentalized into a mitochondrion-derived organelle. While the mitochondrion-related hydrogenosome of *Trichomonas* is a good candidate for this function^{13,14}, a mitochondrion-derived intracellular compartment in *Giardia* has not been identified, despite mounting evidence suggesting its existence^{3–7,15,16}.

We cloned and characterized a second *Giardia* gene involved in the biosynthesis of Fe-S clusters, the *GiiscU* gene. This gene encodes the soluble iron-binding protein IscU, a functional partner of IscS thought to participate as a scaffold protein that forms Fe-S molecular complexes in conjunction with IscS (ref. 11). Amino acid sequence comparison of the *Giardia* IscU with bacterial and eukaryotic homologues revealed a high degree of similarity along the entire length of the protein (63–69% similarity, 46–51% identity), with conservation of cysteine residues essential for *in vivo* functionality¹⁷. Similar to IscS (ref. 5), phylogenetic analyses provide significant support for the clustering of *Giardia* IscU with

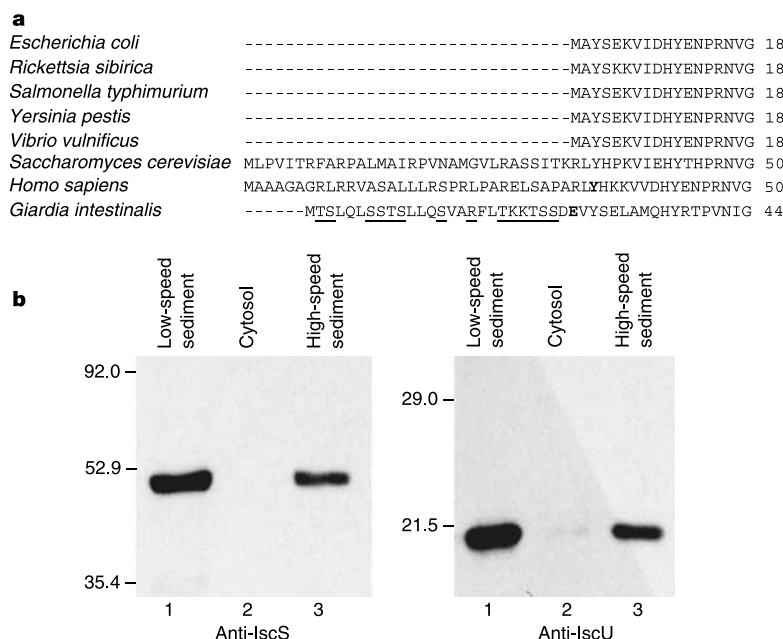


Figure 1 Distribution of IscU and IscS in *Giardia* cell extracts. **a**, Comparison of N-terminal regions of prokaryotic and eukaryotic IscU proteins aligned by CLUSTALW. Hydroxylated and basic amino acids of the *Giardia* sequence are underlined. E27, the proposed N-terminal cleavage site of the *Giardia* sequence and Y35, the first residue of the mature

human IscU (ref. 18), are in bold typeface. **b**, Distribution of Isc proteins in *Giardia* fractions obtained by differential centrifugation. Trophozoite extracts fractionated as described in Methods were electrophoresed, blotted and probed with specific antibodies against IscS and IscU. Sizes of molecular markers are given in kDa.

mitochondrial homologues (not shown). The amino-terminal sequence of IscU is rich in hydroxylated and basic amino acids reminiscent of mitochondrial-targeting presequences (Fig. 1a). Analysis of this sequence using MitoProt predicts the mitochondrial localization of *Giardia* IscU with high confidence ($P > 0.95$). The predicted N-terminal cleavage site at E27 coincides with the first amino acid residue of all prokaryotic IscU sequences and is two residues upstream of Y35, the cleavage site of the human sequence¹⁸. These findings suggest that the N-terminal sequence might target IscU to a mitochondrion-derived organelle in this organism.

To test this hypothesis we overexpressed *Giardia* IscU and IscS in *Escherichia coli* and generated antibodies against the recombinant proteins. The specific anti-IscU and anti-IscS antibodies were used for immunoblotting of subcellular fractions obtained by differential centrifugation. Protein bands of around 49 and 19 kDa were identified, in good agreement with the sizes of IscS and IscU respectively, predicted from their genomic sequences. Both proteins were found predominantly in the low-speed sediment of unbroken cells and cell debris (see Methods for details) and in the high-speed pellet containing sedimentable particulate matter such as membranes and membrane-bounded organelles (Fig. 1b, lanes 1 and 3). Little or no label was found in the cytosol, suggesting that both proteins are compartmentalized in this organism.

We tested the functionality of *Giardia* Fe-S protein maturation using an *in vitro* reconstitution assay that monitors the assembly of *de novo* synthesized radiolabelled Fe-S clusters into recombinant ferredoxin apoprotein lacking Fe-S moieties. Formation of fully assembled ferredoxin occurs exclusively when the high-speed sediment is used in the reaction (Fig. 2). Incubation with cytosolic proteins did not produce significant incorporation of label into ferredoxin and the addition of EDTA, a metal chelator, completely inhibited ferredoxin maturation. These results are consistent with the observed cellular distribution of Isc proteins in *Giardia* and demonstrate that Fe-S cluster biosynthesis, a mitochondrial biosynthetic pathway, is fully operational in this amitochondrial organism.

The compartmentalized distribution of IscS and IscU in *Giardia* was demonstrated by confocal scanning fluorescence microscopy. Fixed trophozoites treated with anti-IscS and anti-IscU antibodies displayed localized staining in small cellular structures distributed throughout the cytoplasm, including an accumulation of label around basal bodies and the base of flagellar axonemes (Fig. 3a; see also Fig. 4f). Figure 3a shows representative maximum projection images for both proteins merged with the corresponding differential interference contrast (DIC) images. The number of organelles detected per cell ranged from 25 to well over 100, as estimated by counting labelled structures in optical stacks of 300 nm thickness, suggesting a significant physiological role for this orga-

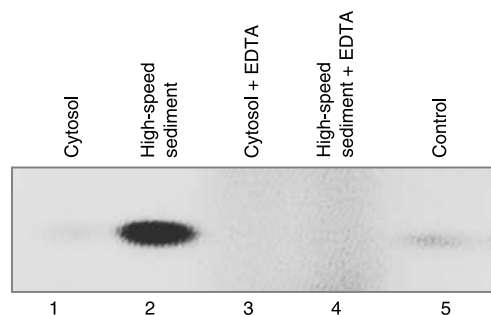


Figure 2 Functional analysis of Fe-S protein maturation in *Giardia*. Incorporation of radioactive Fe-S clusters into recombinant apoferrdoxin was assayed as described in Methods. Assembly of the *de novo* synthesized Fe-S clusters into apoferrdoxin is observed exclusively when the high-speed sediment is used in the reaction. Ferrerdoxin maturation is inhibited by metal ion chelation with EDTA.

nelle in *Giardia*. Double labelling for IscS and IscU revealed that both proteins are present in the same intracellular compartment (Fig. 3b; see also Figs 4c, d), as would be expected from their functional partnership in Fe-S cluster biosynthesis. Interestingly, chaperonin Cpn60 was not found to colocalize with IscS in double labelling experiments (Fig. 3c), indicating that the current function of Cpn60 in *Giardia* is not conducted in the same compartment as Fe-S cluster biosynthesis. Unlike IscU, *Giardia* IscS lacks a recognizable putative mitochondrial targeting signal⁵ but its clear compartmentalized localization suggests that organellar protein import mechanisms in *Giardia* are not solely dependent on N-terminal presequences. A similar observation was recently made in *Trachipleistophora*, where mtHsp70 is efficiently targeted into mito-

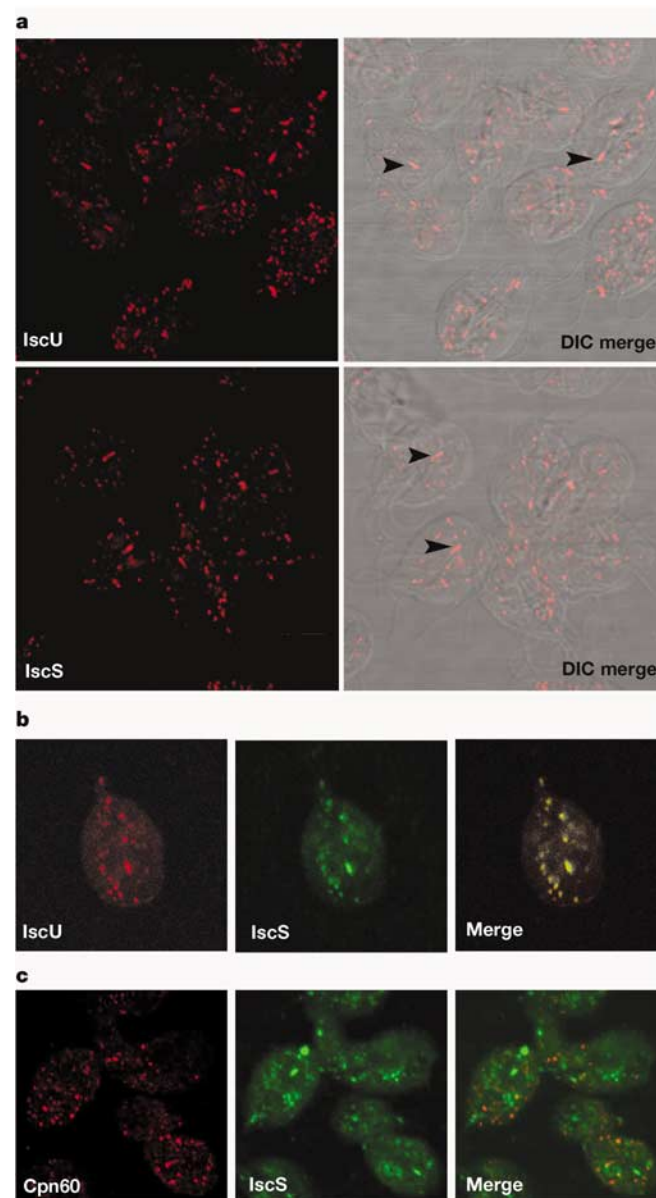


Figure 3 Localization of IscS and IscU in *Giardia* trophozoites by confocal immunofluorescence microscopy. **a**, *Giardia* trophozoites, fixed and permeabilized as described in Methods, were probed independently with antibodies specific for IscU and IscS. The label is located in foci throughout the cytoplasm; distinctive accumulation is also observed around the base of flagellar axonemes and basal bodies (indicated by arrowheads). **b**, Double labelling of *Giardia* trophozoites with anti-IscS and anti-IscU antibodies; colocalization of label is readily apparent. **c**, Double labelling of trophozoites with anti-IscS and anti-Cpn60 antibodies; no colocalization of label is observed.

chondrial remnant organelles despite the absence of a recognizable organelle targeting presequence¹⁰.

We used immunoelectron microscopy to investigate the structure of the organelle harbouring the *Giardia* Isc proteins. Immunogold labelling of IscS and IscU was localized over small organelles bounded by two limiting membranes (Fig. 4a–d). Labelling densities for IscS and IscU were at least 200 and 110 fold higher than over cytosol or peripherally located endolysosomes respectively (Fig. 4i). Double labelling for IscS and IscU localized both of these antigens in the same class of double membrane-bounded structures (Fig. 4c, d), and similar structures were found in epoxy resin sections (Figs 4e–h). In cryosections, gold-labelled double-membrane-bounded organelles were elongated in profile and measured an average of 142×61 nm while in epoxy resin sections these structures measured an average of 137×68 nm (Fig. 4j). These organelles were found distributed throughout the cytosol and in close proximity to flagellar axonemes (Fig. 4f), in agreement with the distribution observed by confocal microscopy.

Immunogold labelling of Cpn60 did not produce significant signal over membrane-bounded structures (not shown), in agreement with a previous report on the lack of association of *Giardia* Cpn60 with biological membranes¹⁶ and consistent with our confocal microscopy observations. The differential distribution of Cpn60 and IscS is noteworthy because the mitochondrial ancestry of both proteins is strongly supported by phylogenetic analyses.

However, the fact that only Isc proteins are associated with membrane-bounded organelles suggests that Cpn60 has been relocated from its original mitochondrial compartment as part of an ongoing process of reductive evolution. Alternatively, the possibility that *Giardia* Cpn60 might have been acquired by lateral transfer from a prokaryotic organism related to the mitochondrial endosymbiont cannot be excluded^{19,20}. The precise nature of the Cpn60 aggregates that render the characteristic punctate distribution in *Giardia* remains to be elucidated.

Overall, our data provide direct physical evidence for the presence of mitochondrial remnant organelles (mitosomes) in *Giardia*, as postulated on the basis of previous phylogenetic and biochemical studies^{3–7,15,21}. Mitochondrial remnant organelles have also been observed in other amitochondrial protists including the microsporidian *Trachipleistophora hominis*, the apicomplexan *Cryptosporidium parvum* and the entamoebid *Entamoeba histolytica*^{8–10,22}. *Giardia* has traditionally been considered to be one of the earliest-branching eukaryotes, although the significance of its basal position in phylogenetic reconstructions has been disputed^{1,2,23}. The identification of mitosomes in this organism places the endosymbiotic event that led to mitochondria firmly before the evolutionary divergence of the diplomonads.

Energy generation and recycling of resources are thought to have driven the original endosymbiotic event. Whether ATP generation in the heterotrophic mitochondrial endosymbiont involved aerobic

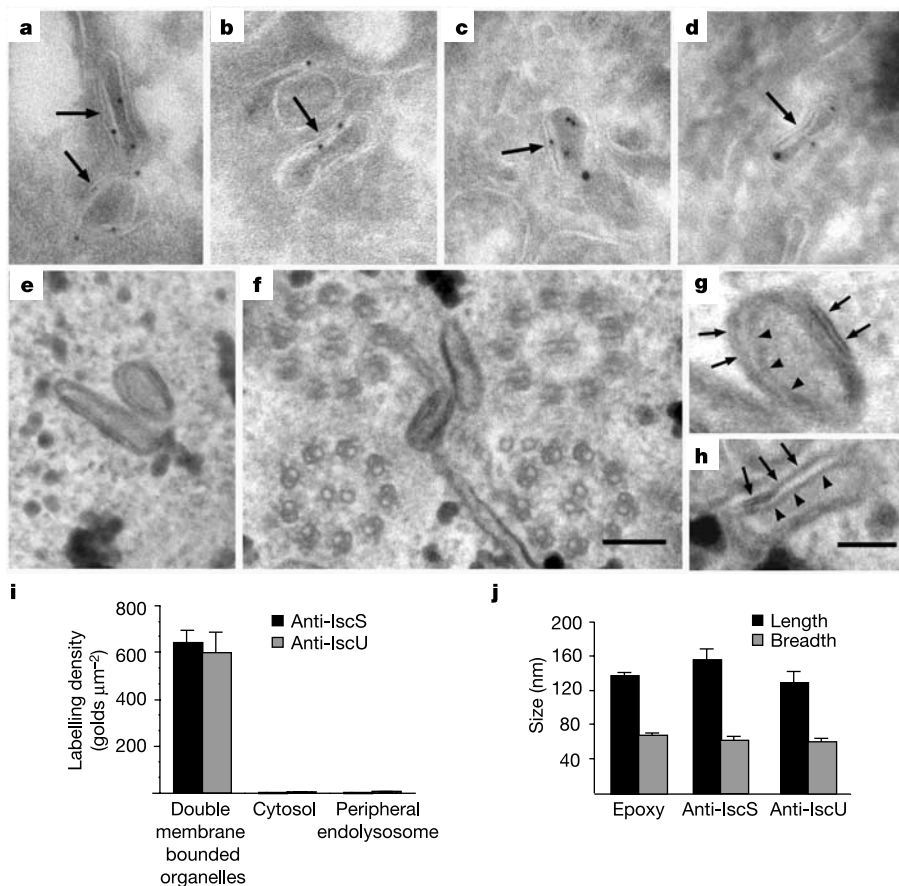


Figure 4 Localization of IscS and IscU in *Giardia* trophozoites by transmission electron immunomicroscopy. **a–d**, Thawed frozen cryosections of glutaraldehyde-fixed *Giardia* were labelled with anti-IscS (**a**) and anti-IscU antibodies (**b**) or double-labelled (**c** and **d**) for anti-IscU (8 nm gold particles) followed by anti-IscS antibodies (12 nm gold particles). Arrows indicate double-membrane-bounded structures. Membranes appear as white profiles with this methodology. **e–h**, Epoxy resin sections of glutaraldehyde-fixed trophozoites. **f**, Double-membrane-bounded structures are shown in close proximity to flagellar axonemes. **g, h**, High-magnification images showing organelle double

membranes. Arrows point towards outer membranes while arrowheads identify inner membranes. **i**, Densities of labelling from typical labelling experiments were obtained as explained in Methods ($n = 8$ for IscS and $n = 7$ for IscU; $n =$ number of scanned micrographs). **j**, Sizes of double membrane bounded organelles in epoxy resin and cryosections labelled with antibodies to either IscS or IscU ($n = 43$ for epoxy resin, $n = 13$ for IscS and $n = 11$ for IscU; $n =$ number of organelles). Intermembrane spacing in epoxy resin was 9.36 ± 0.49 nm ($n = 29$). Scale bars, 100 nm (**a–f**) and 50 nm (**g, h**).

or anaerobic respiration—as proposed by current theories of eukaryotic evolution (reviewed in ref. 24)—all electron transport chains and modes of heterotrophic oxidative metabolism share a common feature: several of their key enzymes rely on Fe–S centres for their catalytic activities (for example, ferredoxins, pyruvate:ferredoxin oxidoreductase, succinate dehydrogenase, cytochrome *b*-associated Rieske protein, aconitase, hydrogenase)¹². Thus, the original endosymbiont must have possessed the capacity to synthesize Fe–S clusters and to assemble them into functional redox and electron transport proteins. These crucial metabolic functions could have secured the permanent establishment of the mitochondrial endosymbiont in the newly formed eukaryotic cell.

Our data also directly demonstrate an essential metabolic function for the mitochondria of amitochondriate protists. The apparent ubiquity of Fe–S cluster assembly proteins in the genomes of all amitochondriate protists sequenced until now (for example, *Encephalitozoon*²⁵, *Giardia*²⁶, *Cryptosporidium*²⁷, *Entamoeba* (http://www.sanger.ac.uk/Projects/E_histolytica/), (<http://www.tigr.org/tdb/e2k1/eha1/>)) suggests that Fe–S cluster biosynthesis and Fe–S protein maturation might have been the only biochemical features of the mitochondrial endosymbiont essential for eukaryotic cell survival under conditions of oxygen deprivation. Interestingly, hydrogen production by *Giardia* has recently been detected under strict anaerobic conditions²⁸. The possibility that this process might also occur in the double-membrane-bounded organelles described here awaits the unequivocal localization of hydrogenase in this organism. We suggest that the requirement for the assembly and maturation of Fe–S proteins may have provided, for hundreds of millions of years, the selective pressure required for the retention of the mitochondrial endosymbiont in both aerobic and anaerobic environments. □

Methods

Gene cloning, protein expression and antibody generation

The entire coding regions of the *Giardia* *iscS* and *iscU* genes were amplified by PCR using *Giardia* genomic DNA as template and the primer pairs 5′-TATTTTAAATGAGGTACC TGGACAACCAAGCCACTACG-3′ / 5′-AGTGGTGTGAGTCGACTAGTCATGCTT CCACTCTATGCTC-3′ and 5′-AATTCTAATGGATCCCAAGCCTTCAGCTCTCTA GCACT-3′ / 5′-CAGGAGGCACCCGGGTCAAGAAGACTTTGATACCTGTATC-3′ designed on the basis of partial sequences retrieved from the *Giardia* genome database²⁶. These incorporated unique *KpnI* and *SalI* sites and *BamHI* and *XmaI* sites (italics) into *iscS* and *iscU* respectively. PCR products were cloned directionally into plasmid pQE-32 (Qiagen) and transformed into *E. coli* M15 [pREP4]. N-terminal His-tagged proteins were purified under denaturing conditions by affinity chromatography following the manufacturer's protocol (Qiagen) and used to generate rabbit polyclonal anti-*iscS* and anti-*iscU* antibodies (Lampire Biological Laboratories). Mouse polyclonal anti-Cpn60 antibodies were generated using recombinant Cpn60 protein from *E. histolytica* prepared as described⁹.

Organism, cell fractionation and western blotting

Giardia intestinalis (WB strain, ATCC No. 30957) was used throughout this study. Trophozoites were partially lysed to a maximum 50% cell lysis in isotonic buffer (0.2 M mannitol, 50 mM sucrose, 10 mM KCl, 1 mM EDTA, 10 mM HEPES pH 7.4 supplemented with 70 μM E64 and 0.75 μg ml⁻¹ leupeptin). A low-speed sediment fraction composed of cell debris and unbroken cells was obtained by centrifugation at 1,000g for 10 min at 4 °C. The supernatant was further centrifuged at 108,000g for 30 min. The resulting high-speed sediment was resuspended in isotonic buffer. The high-speed supernatant was used as the soluble cytosolic fraction. Equivalent protein aliquots (10 μg) were electrophoresed on 10% or 12.5% polyacrylamide gels for detection of *iscS* and *iscU* respectively. Proteins were electroblotted onto nitrocellulose membranes, blocked by incubation in 2% skimmed milk in PBS and probed with anti-*iscS* (1:1,000) or anti-*iscU* (1:750) antibodies for 1 h at room temperature. Membranes were further incubated with anti-rabbit IgG antibodies conjugated to peroxidase (1:25,000) for 1 h at room temperature. Blots were developed using an ECL detection system (Amersham Biosciences).

Reconstitution of Fe–S clusters *in vitro*

Trophozoites were mechanically disrupted and fractionated as described above. Fe–S cluster formation by cytosolic and high-speed sediment fractions was determined using recombinant apoferritin from *Trichomonas vaginalis* as a reporter protein. Recombinant ferredoxin was prepared as described²⁹. The standard reaction mixture contained 30–100 μg protein of tested fraction, 7 μg of apo-ferredoxin, 20 mM HEPES, 0.2% Triton-X100, 250 mM saccharose, 50 μM ferrous ascorbate, 20 mM DTT, 25 μM L-cysteine and 10 μCi of ³⁵S-L-cysteine. To observe inhibition of the reaction, 5 mM EDTA was added to the reaction mixture before the incubation. In controls, the cell fractions

were omitted. Reactions proceeded in an anaerobic jar under atmosphere of 95% nitrogen and 5% hydrogen gas for 90 min at 25 °C. Samples were separated on a 15% non-denaturing polyacrylamide gel at 4 °C, followed by gel vacuum-drying and autoradiography.

Immunofluorescence confocal microscopy

Giardia trophozoites were fixed in 4% paraformaldehyde in PBS for 30 min at 37 °C and permeabilized with cold acetone or 0.2% Triton X-100 in PBS for 5 min. After blocking with 1% BSA in PBS, slides were incubated overnight with anti-*iscS* rabbit antibodies (1:200 dilution) and subsequently with secondary anti-rabbit IgG-TRITC conjugate (1:300 dilution) for 1 h at 37 °C. In *iscS* colocalization experiments, slides treated with anti-*iscU* antibodies were further incubated with 250 μl of purified rabbit IgG immunoglobulins (40 μg ml⁻¹) for 1 h at 37 °C followed by incubation with FITC-conjugated anti-*iscS* antibodies (Fluoro Tag FITC Conjugation Kit, Sigma). Slides were mounted with anti-fade mounting medium (Vectashield) and observed under a laser scanning confocal microscope (Radiance 2100, Bio-Rad). Images were collected using Bio-Rad LaserSharp 2000 software and processed with Adobe Photoshop 7.0.

Immunoelectron microscopy

Organisms were fixed in 0.5% glutaraldehyde or 4% paraformaldehyde/0.1% glutaraldehyde each in 0.2 M PIPES buffer pH 7.2 for 30 min at room temperature, washed in PBS and stored at 4 °C. Cells were pelleted and cryoprotected in 2.3 M sucrose in PBS and frozen on cryosectioning stubs in liquid nitrogen and cryosectioned on a Leica ultracycromicrotome at 50–100 nm. Thawed cryosections were labelled using rabbit antisera against *iscS* proteins followed by 8 nm protein A gold³⁰, contrasted in methyl cellulose/uranyl acetate and observed in a JEOL 1200EX transmission electron microscope. For quantitation, elongated or round profiles with evidence of double membranes and measuring less than 110 nm minor diameters were photographed at a nominal magnification of 40,000. Micrographs were recorded on phosphorimaging plates (4,342 × 4,912 pixels) (DITABIS, Germany) and stored as TIFF files and further magnified in Adobe Photoshop 5.5. Square lattice grids with spacing of 25 nm, 250 nm and 125 nm were generated on screen and point counting was used to estimate areas of double membrane structures, cytosol and peripherally located endolysosomes respectively. Double labelling using 8 nm and 12 nm gold particles was carried out as described elsewhere³⁰. Standard errors of ratio estimates were computed according to ref. 31.

Received 7 April; accepted 22 July 2003; doi:10.1038/nature01945.

- Adam, R. D. Biology of *Giardia lamblia*. *Clin. Microbiol. Rev.* **14**, 447–475 (2001).
- Sogin, M. L. & Silberman, J. D. Evolution of the protists and protistan parasites from the perspective of molecular systematics. *Int. J. Parasitol.* **28**, 11–20 (1998).
- Hashimoto, T., Sánchez, L. B., Shirakura, T., Müller, M. & Hasegawa, M. Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. *Proc. Natl Acad. Sci. USA* **95**, 6860–6865 (1998).
- Roger, A. J. *et al.* A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. *Proc. Natl Acad. Sci. USA* **95**, 229–234 (1998).
- Tachezy, J., Sánchez, L. B. & Müller, M. Mitochondrial type iron-sulfur cluster assembly in the amitochondriate eukaryotes *Trichomonas vaginalis* and *Giardia intestinalis*, as indicated by the phylogeny of *iscS*. *Mol. Biol. Evol.* **18**, 1919–1928 (2001).
- Morrison, H. G., Roger, A. J., Nystul, T. G., Gillin, F. D. & Sogin, M. L. *Giardia lamblia* expresses a proteobacterial-like DnaK homolog. *Mol. Biol. Evol.* **18**, 530–541 (2001).
- Arisue, N., Sánchez, L. B., Weiss, L. M., Müller, M. & Hashimoto, T. Mitochondrial-type hsp70 genes of the amitochondriate protists *Giardia intestinalis*, *Entamoeba histolytica* and two microsporidians. *Parasitol. Int.* **51**, 9–16 (2002).
- Mai, Z. *et al.* Hsp60 is targeted to a cryptic mitochondrion-derived organelle (“crypton”) in the microaerophilic protozoan parasite *Entamoeba histolytica*. *Mol. Cell. Biol.* **19**, 2198–2205 (1999).
- Tovar, J., Fischer, A. & Clark, C. G. The mitochondrion, a novel organelle related to mitochondria in the amitochondrial parasite *Entamoeba histolytica*. *Mol. Microbiol.* **32**, 1013–1021 (1999).
- Williams, B. A. P., Hirt, R. P., Lucocq, J. M. & Embley, T. M. A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. *Nature* **418**, 865–869 (2002).
- Lill, R. & Kispal, G. Maturation of cellular Fe-S proteins: an essential function of mitochondria. *Trends Biochem. Sci.* **25**, 352–356 (2000).
- Müller, M. in *Molecular Medical Parasitology* (eds Marr, J., Nilsen, T. & Komuniecki, R.) 125–139 (Academic, London, 2003).
- Dyall, S. D. & Johnson, P. J. Origins of hydrogenosomes and mitochondria: evolution and organelle biogenesis. *Curr. Opin. Microbiol.* **3**, 404–411 (2000).
- Embley, T. M., van der Giezen, M., Horner, D. S., Dyall, P. L. & Foster, P. Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Phil. Trans. R. Soc. Lond.* **358**, 191–203 (2003).
- Lloyd, D. & Harris, J. C. *Giardia*: highly evolved parasite or early branching eukaryote? *Trends Microbiol.* **10**, 122–127 (2002).
- Soltys, B. J. & Gupta, R. S. Presence and cellular distribution of a 60-kDa protein related to mitochondrial hsp60 in *Giardia lamblia*. *J. Parasitol.* **80**, 580–590 (1994).
- Tokumoto, U. *et al.* Network of protein-protein interactions among iron-sulfur cluster assembly proteins in *Escherichia coli*. *J. Biochem. (Tokyo)* **131**, 713–719 (2002).
- Tong, W. H. & Rouault, T. Distinct iron-sulfur cluster assembly complexes exist in the cytosol and mitochondria of human cells. *EMBO J.* **19**, 5692–5700 (2000).
- Andersson, J. O., Sjogren, A. M., Davis, L. A., Embley, T. M. & Roger, A. J. Phylogenetic analyses of diplomonads genes reveal frequent lateral gene transfers affecting eukaryotes. *Curr. Biol.* **13**, 94–104 (2003).
- Doolittle, W. F. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* **14**, 307–311 (1998).
- Lloyd, D. *et al.* The ‘primitive’ microaerophile *Giardia intestinalis* (syn. *lamblia*, *duodenalis*) has

- specialized membranes with electron transport and membrane-potential-generating functions. *Microbiology* **148**, 1349–1354 (2002).
22. Riordan, C. E., Ault, J. G., Langreth, S. G. & Keithly, J. S. *Cryptosporidium parvum* Cpn60 targets a relict organelle. *Curr. Genet. advance online publication*, 20 August 2003 (doi:10.1007/s00294-003-0432-1).
 23. Philippe, H. *et al.* Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc. R. Soc. Lond. B* **267**, 1213–1221 (2000).
 24. Martin, W., Hoffmeister, M., Rotte, C. & Henze, K. An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. *Biol. Chem.* **382**, 1521–1539 (2001).
 25. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**, 450–453 (2001).
 26. McArthur, A. G. *et al.* The *Giardia* genome project database. *FEMS Microbiol. Lett.* **189**, 271–273 (2000).
 27. Strong, W. B. & Nelson, R. G. Preliminary profile of the *Cryptosporidium parvum* genome: an expressed sequence tag and genome survey sequence analysis. *Mol. Biochem. Parasitol.* **107**, 1–32 (2000).
 28. Lloyd, D., Ralphs, J. R. & Harris, J. C. *Giardia intestinalis*, a eukaryote without hydrogenosomes, produces hydrogen. *Microbiology* **148**, 727–733 (2002).
 29. Vidakovic, M. S., Fraczekiewicz, G. & Germanas, J. P. Expression and spectroscopic characterization of the hydrogenosomal [2Fe-2S] ferredoxin from the protozoan *Trichomonas vaginalis*. *J. Biol. Chem.* **271**, 14734–14739 (1996).
 30. Prescott, A. R., Lucocq, J. M., James, J. M. & Ponnambalam, S. Distinct compartmentalization of TGN46 and beta 1,4-galactosyl transferase in HeLa cells. *Eur. J. Cell Biol.* **72**, 238–246 (1997).
 31. Cochran, W. G. *Sampling Techniques* (John Wiley and Sons, London, 1977).

Acknowledgements We thank G. Clark, J. Bowyer and S. Cutting for critically reading the manuscript. A recombinant plasmid containing the *T. vaginalis* ferredoxin gene was provided by J. P. Germanas and K. Krause. The use of partial genome sequence information from the *Giardia* Genome Project Database²⁶ is acknowledged. The technical assistance of J. James and N. Sommerville is also acknowledged. M.H. is a sabbatical visitor supported by CINVESTAV, México. Research at the Rockefeller University (gene cloning, antibody generation) was supported by a NIH grant to M.M. Research at Charles University (*in vitro* assembly of Fe-S clusters) was supported by a grant from FIRCA to J. Tachezy. J.M.L. (electron microscopy) was supported by a Research Leave Fellowship from the Wellcome Trust and by Tenovus Scotland. Research at Royal Holloway (bioinformatics, cell fractionation, fluorescence confocal microscopy, manuscript writing, project coordination) was supported by a Wellcome Trust grant to J. Tovar.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J. Tovar (j.tovar@rhul.ac.uk). The sequence of *GiUC* has been deposited with GenBank under accession number AY040612.

Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers

H. D. Bradshaw Jr¹ & Douglas W. Schemske²

¹Department of Biology, University of Washington, Seattle, Washington 98195, USA

²Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824, and W. K. Kellogg Biological Station, Hickory Corners, Michigan 49060, USA

The role of major mutations in adaptive evolution has been debated for more than a century^{1,2}. The classical view is that adaptive mutations are nearly infinite in number with infinitesimally small phenotypic effect³, but recent theory suggests otherwise⁴. To provide empirical estimates of the magnitude of adaptive mutations in wild plants, we conducted field studies to determine the adaptive value of alternative alleles at a single locus, *YELLOW UPPER*^{5–7} (*YUP*). *YUP* controls the presence or absence of yellow carotenoid pigments in the petals of pink-flowered *Mimulus lewisii*, which is pollinated by bumblebees^{5,8–10}, and its red-flowered sister species¹¹ *M. cardinalis*, which is pollinated by hummingbirds^{5,8–10}. We bred near-isogenic lines (NILs) in which the *YUP* allele from each species was substituted into the

other. *M. cardinalis* NILs with the *M. lewisii* *YUP* allele had dark pink flowers and received 74-fold more bee visits than the wild type, whereas *M. lewisii* NILs with the *M. cardinalis* *yup* allele had yellow-orange flowers and received 68-fold more hummingbird visits than the wild type. These results indicate that an adaptive shift in pollinator preference may be initiated by a single major mutation.

Where their ranges overlap, the monkeyflowers *M. lewisii* and *M. cardinalis* are >99% reproductively isolated by the difference in their pollinator guilds^{8,9}. In previous studies of artificial F₂ hybrids between *M. lewisii* and *M. cardinalis*, we showed that flower colour has marked effects on pollinator visitation⁸, and that yellow pigment concentration is controlled in part by the major quantitative trait locus (QTL; reviewed in ref. 12), *YUP*^{6,7}. Although F₂ populations are useful for mapping QTLs controlling differences in floral traits between species^{6,7,13}, they are less than ideal for assessing the adaptive effect of a single mutation. The many intermediate flower phenotypes in an F₂ population^{6–8,13} may provide a bridge for pollinators to develop learned visitation patterns completely unlike those that would occur as a result of a single-locus mutational step in an adaptive walk.

By substituting one allele for another using repeated backcrosses, NILs more closely mimic the effect of a single mutation likely to be part of an adaptive pollinator shift; that is, from bumblebee-pollinated to hummingbird-pollinated, or vice versa. The dominant *M. lewisii* *YUP* allele prevents carotenoid deposition, so the petals show only their pink anthocyanin pigments. The recessive *M. cardinalis* *yup* allele allows carotenoid deposition in the petals and produces red flowers when present in conjunction with a high concentration of anthocyanins^{5–7}. Although phylogenetic evidence suggests that the hummingbird pollination syndrome of *M. cardinalis* is derived from a bee-pollinated ancestor similar to *M. lewisii*¹¹, we constructed *YUP* NILs in both species (Fig. 1). The wild-type *M. lewisii* NIL is pink-flowered (Fig. 1a), whereas the ‘mutant’ NIL homozygous for the introgressed *M. cardinalis* *yup* allele has pale yellow-orange flowers (Fig. 1b). The wild-type *M. cardinalis* NIL is red-flowered (Fig. 1c), but the presence of a dominant *M. lewisii* *YUP* allele produces a dark-pink-flowered NIL (Fig. 1d).

Pollinator visitation rates were determined by field observation of NIL experimental arrays near a zone of sympatry between *M. lewisii* and *M. cardinalis*^{5,8} to ensure that pollinators were familiar with both species in their natural habitat. Bumblebees strongly prefer pink-flowered NILs carrying the *YUP* allele (Fig. 1a, d) in both the *M. lewisii* and *M. cardinalis* genetic backgrounds (Table 1). Hummingbirds prefer yellow-orange- or red-flowered NILs homozygous for the *yup* allele (Fig. 1b, c) in both backgrounds (Table 1).

The striking effect of flower colour on pollinator specificity is evidence for the adaptation of both monkeyflower species to their current pollinators (Table 1). A wild-type pink *M. lewisii* flower (Fig. 1a) is >700 times more likely to be visited by a bumblebee than by a hummingbird, whereas the yellow-orange-flowered ‘mutant’ (Fig. 1b) is only 1.8 times as likely to be visited by a bumblebee. In the *M. cardinalis* background, a wild-type red flower (Fig. 1c) is >1,200 times more likely to be visited by a hummingbird than by a bumblebee, but the pink-flowered ‘mutant’ (Fig. 1d) is visited only 15 times as frequently by hummingbirds.

When these visitation rates are compared with the results from our previous F₂ QTL mapping population⁸, we find that the F₂ experiments accurately predict pollinator visitation when we consider only bumblebees visiting *M. lewisii* NILs, and hummingbirds visiting *M. cardinalis* NILs. In *M. lewisii* NILs and the F₂, the wild-type pink flowers were visited by bumblebees at about a fivefold higher rate than were the ‘mutant’ yellow-orange flowers (Table 1 and ref. 8). In *M. cardinalis* NILs, hummingbirds showed a slight 1.1-fold preference for wild-type red flowers over the pink-flowered ‘mutants’, similar to that found in the F₂ population (Table 1 and ref. 8). The close correspondence of the results from these indepen-



Figure 1 Near-isogenic lines of *M. lewisii* and *M. cardinalis* with alternate alleles at the *YUP* locus. **a, b**, *M. lewisii*; **c, d**, *M. cardinalis*. The wild-type allele at the *YUP* locus (**a, c**) has been substituted by introgression with the allele from the other species (**b, d**). Flowers in each NIL pair (**a** and **b**, **c** and **d**) are full siblings.

dent experiments suggests that they address the same question: what is the effect of a *YUP* mutation on visitation by the current pollinator?

From an evolutionary perspective, it is perhaps more illuminating to ask a different question: what is the effect of a *YUP* mutation on the attraction of a novel pollinator guild, as would be the relevant scenario for a new mutation leading to an adaptive shift from one pollinator guild to another in the common ancestor of *M. lewisii* and *M. cardinalis*? Our NIL experiments reveal that hummingbirds visit yellow-orange-flowered 'mutants' of *M. lewisii* at 68 times the rate of the pink-flowered wild type, and bumblebees visit pink-flowered 'mutants' of *M. cardinalis* at 74 times the rate of the red-flowered wild type (Table 1). The large and symmetrical effect of the *YUP* allele substitution on the attraction of a new pollinator guild implies that a mutation at the *YUP* locus has the potential to alter the pollinator assemblage dramatically in the common ancestor of *M. lewisii* and *M. cardinalis*.

As 'mutations' at the *YUP* locus decrease visitation by the current pollinator guild, and simultaneously increase visitation by a new pollinator guild, are there plausible ecological circumstances in which the mutant might be favoured by natural selection? The combined rate of bumblebee and hummingbird visitation to the yellow-orange-flowered 'mutants' of *M. lewisii* is just 26% of that to the wild-type pink flowers, and the combined rate for dark-pink-flowered 'mutants' of *M. cardinalis* is 95% of the wild type. This implies that a change in the relative abundance of bumblebees and hummingbirds, compared with the pollinator assemblage present during our field experiments, would be required for the mutant to be favoured by natural selection in the common ancestor of

Table 1 Pollinator visitation rates to NILs of *M. lewisii* and *M. cardinalis*

	Bumblebees (10^{-3} visits per flower per hour)	Hummingbirds (10^{-3} visits per flower per hour)
<i>M. lewisii</i> NILs		
Wild-type (pink; Fig. 1a)	15.4	0.0212
'Mutant' (yellow-orange; Fig. 1b)	2.63	1.44
<i>M. cardinalis</i> NILs		
Wild-type (red; Fig. 1c)	0.148	189
'Mutant' (dark pink; Fig. 1d)	10.9	168

Note that the visitation rates estimated for bumblebees to red-flowered *M. cardinalis* NILs and for hummingbird visits to pink-flowered *M. lewisii* NILs are likely to be less accurate owing to the small absolute number of visits ($N = 2$ and $N = 1$, respectively).

M. lewisii and *M. cardinalis*. The change in relative abundance of pollinators necessary to produce equal visitation to both flower colour phenotypes can be estimated from our data. A ninefold decrease in the relative abundance of bumblebees would produce equal combined visitation rates in the wild-type pink-flowered and 'mutant' yellow-orange-flowered *M. lewisii* NILs. At the equilibrium point, 99% of visitors to wild-type *M. lewisii* flowers would be bumblebees, whereas 87% of visitors to 'mutants' would be hummingbirds. In the *M. cardinalis* NILs, a twofold increase in the relative abundance of bumblebees would produce equal visitation rates to pink and red flowers. At the equilibrium point, hummingbirds would be virtually the only visitor to the wild-type red *M. cardinalis* flowers, and remain the major visitor (89% of visits) even to the dark-pink 'mutants'.

The evolution of hummingbird-pollinated flowers from insect-pollinated ancestors is a recurring theme in the flora of western North America¹⁴. A molecular phylogenetic analysis of *Mimulus* indicates that hummingbird pollination has evolved independently twice within the section *Erythranthe*, in one of these cases leading to the evolution of *M. cardinalis* from an insect-pollinated ancestor likely to have resembled the extant *M. lewisii*¹¹. We have shown that an adaptive divergence in pollinator preference, as might be expected at the speciation event that occurred in the common ancestor of *M. lewisii* and *M. cardinalis*, could in principle be initiated by a single mutation with a large effect on flower colour.

To understand in greater detail the dynamics of an adaptive pollinator shift, it will be necessary to more closely replicate the appearance of a new mutation. First, it must be demonstrated that the recessive allele at the *YUP* locus can be produced by a single loss-of-function mutation, ruling out the possibility that the *YUP* locus contains more than a single gene. Second, a null mutant at the *YUP* locus in *M. lewisii* could be established at a realistic (that is, low) frequency in a natural setting, and its evolutionary trajectory observed. In addition, NILs could be developed carrying 1, 2,...*N* allele substitutions at major QTLs, in various combinations, to test alternative hypotheses for the trajectory of floral evolution and speciation in response to pollinator choice. □

Methods

NIL construction

Near-isogenic lines were derived from two backcross (BC) populations: *M. lewisii* × (*M. lewisii* × *M. cardinalis*) and *M. cardinalis* × (*M. lewisii* × *M. cardinalis*). All NILs were produced by single-seed descent. Ten first-generation backcross (BC₁) plants with *M. lewisii* as the recurrent parent were chosen as the founders of NILs on the basis of their inheritance of a dominant random amplified polymorphic DNA (RAPD) marker (AG13_108; ref. 6) linked in coupling to the recessive *yup* allele (H.D.B. and D.W.S., unpublished work). A single plant from each of these ten *M. lewisii* NILs was backcrossed to a series of unrelated *M. lewisii* recurrent parents for three additional generations, maintaining selection for the AG13_108 marker. After four generations of backcrossing, each NIL is expected to share 97% of its genome with the recurrent parent. For each of the ten lines, a single BC₄ plant was self-pollinated to produce ten BC₄S₁ families segregating at the *YUP* locus. The BC₄S₁ families yielded both wild-type pink and 'mutant' yellow-orange flowers. Ten *M. cardinalis* NILs were constructed using the same mating design, except that selection for the presence of the dominant *YUP* allele was done visually (dark-pink flowers), and self-pollination of the BC₄ generation was unnecessary because each generation segregated 1:1 for red: pink flowers.

Recurrent parent similarity index

Six characters (upper petal reflexing, lateral petal reflexing, pistil length, stamen length, lateral petal width and nectar volume) for which QTLs have been mapped^{6,7} were measured on two flowers from each plant. There was a significant difference between the multivariate flower phenotypes of wild-type and 'mutant' NILs in both the *M. lewisii* (multiple analysis of variance, MANOVA, $F = 18.18$, Wilks' $\lambda = 0.32$, $P < 0.0001$) and *M. cardinalis* (MANOVA, $F = 11.00$, Wilks' $\lambda = 0.56$, $P < 0.0001$) genetic backgrounds (PROC GLM, SAS Institute). Least-squares means for each trait within each NIL genotypic class were normalized to the difference between trait means of the two parental species⁷, setting the recurrent parent trait value at 100% and the nonrecurrent parent at 0%. Lower recurrent parent similarity (RPS) values are evidence of linkage drag, whereas values larger than 100% represent measurement error or heterosis. In the *M. lewisii* genetic background, the wild-type plants had a mean RPS index across all traits of 91% (range 66–108%), whereas their 'mutant' sibs had a value of 80% (range 51–103%). In the *M. cardinalis* genetic background, the wild-type plants had a mean RPS index of 95% (range 49–129%), whereas their 'mutant' sibs had a value of 80% (range 46–155%). Although 'mutant' NILs show more linkage drag than the wild type, we judge the difference to be small. Nectar volume, which is known from our F_2 experiments to have a marked effect on hummingbird visitation⁸, has RPS index values that are very close to one another in the NILs: 105% and 103% in the *M. lewisii* background, and 46% and 53% in the *M. cardinalis* background. This suggests that differences in nectar production between pairs of NILs did not affect pollinator visitation patterns.

Pollinator visitation

For each of two field experiments conducted to measure pollinator visitation, 50 pink or dark pink (*YUP*____) and 50 yellow-orange or red (*yup/yup*) plants were drawn at random from five BC₄S₁ (*M. lewisii*) or BC₄ (*M. cardinalis*) NIL families. Assessments of pollinator visitation were performed at Mather (California, USA), the site where much of the previous work on these two species of *Mimulus* has been done⁵. Pollinator observations were carried out from dawn to evening, with a 1–2 h break at midday when pollinators were least active. Dates of observation were 18–30 August 1999 for *M. cardinalis* NILs, and 18–27 July 2000 for *M. lewisii* NILs. These dates correspond closely to the peak flowering times of natural populations of the two *Mimulus* species. We chose to do the experiments in different years so that pollinators were faced with a binary choice of flower phenotypes, as would be the case for a newly arisen mutation. Plants were placed at random on a 1 m × 1 m grid to produce the experimental arrays (a black bear visit reduced the total sample size in the *M. lewisii* NIL array from $N = 100$ to $N = 99$). A pollinator visit was counted if it appeared that the pollinator probed the flower and contacted the anthers or stigma. Bumblebees and hummingbirds were the only pollinators observed. We observed 1,090 bumblebee visits to the *M. lewisii* NILs, 180 bumblebee visits to the *M. cardinalis* NILs, 201 hummingbird visits to the *M. lewisii* NILs, and 3,738 hummingbird visits to the *M. cardinalis* NILs. The number of flowers on each plant was recorded daily, along with the number of hours spent observing. Visitation rates were calculated by dividing the total number of pollinator visits across all days by the aggregate number of hours in which visits could have occurred to each flower (flower-hours). For the *M. lewisii* NILs, both bumblebee and hummingbird pollinator observations were carried out simultaneously, with 47,159 flower-hours for the wild-type NILs and 138,648 flower-hours for the 'mutants'. For the *M. cardinalis* NILs, separate pollinator observation periods were required to keep track of the large number of hummingbird visits. During the bumblebee observation periods, there were 16,291 flower-hours for the 'mutant' NILs and 13,556 flower-hours for the wild-type. During the hummingbird observation periods, there were 11,505 flower-hours for the 'mutant' NILs and 9,520 flower-hours for the wild type.

Received 15 July; accepted 3 October 2003; doi:10.1038/nature02106.

- Orr, H. A. & Coyne, J. A. The genetics of adaptation: a reassessment. *Am. Nat.* **140**, 725–742 (1992).
- Gillham, N. W. Evolution by jumps: Francis Galton and William Bateson and the mechanism of evolutionary change. *Genetics* **159**, 1383–1392 (2001).
- Fisher, R. A. *The Genetical Theory of Natural Selection* (Dover, New York, 1958).
- Orr, H. A. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* **52**, 935–949 (1998).
- Hiesey, W. M., Nobs, M. A. & Björkman, O. *Experimental Studies on the Nature of Species: V. Biosystematics, Genetics, and Physiological Ecology of the Erythranthe Section of Mimulus* **16** (Carnegie Inst. Wash. Publ. 628, Washington DC, 1971).
- Bradshaw, H. D. Jr, Wilbert, S. M., Otto, K. G. & Schemske, D. W. Genetic mapping of floral traits associated with reproductive isolation in monkeyflowers (*Mimulus*). *Nature* **376**, 762–765 (1995).
- Bradshaw, H. D. Jr, Otto, K. G., Frewen, B. E., McKay, J. K. & Schemske, D. W. Quantitative trait loci affecting differences in floral morphology between two species of monkeyflower (*Mimulus*). *Genetics* **149**, 367–382 (1998).
- Schemske, D. W. & Bradshaw, H. D. Jr Pollinator preference and the evolution of floral traits in monkeyflowers (*Mimulus*). *Proc. Natl Acad. Sci. USA* **96**, 11910–11915 (1999).
- Ramsey, J., Bradshaw, H. D. Jr & Schemske, D. W. Components of reproductive isolation between the monkeyflowers *Mimulus lewisii* and *M. cardinalis* (Phrymaceae). *Evolution* **57**, 1520–1534 (2003).
- Vickery, R. K. Jr Speciation in *Mimulus*, or, can a simple flower color mutant lead to species divergence? *Great Basin Nat.* **55**, 177–180 (1995).
- Beardsley, P. M., Yen, A. & Olmstead, R. G. AFLP phylogeny of *Mimulus* section Erythranthe and the evolution of hummingbird pollination. *Evolution* **57**, 1397–1410 (2003).
- Mauricio, R. Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nature Rev. Genet.* **2**, 370–381 (2001).
- Hodges, S. A., Whittall, J. B., Fulton, M. & Yang, J.-Y. Genetics of floral traits influencing reproductive isolation between *Aquilegia formosa* and *Aquilegia pubescens*. *Am. Nat.* **159**, S51–S60 (2002).
- Grant, V. Historical development of ornithophily in the western North American flora. *Proc. Natl Acad. Sci. USA* **91**, 10407–10411 (1994).

Acknowledgements We thank A. Angert, K. Kay, and D. Grosenbacher for field observations of pollinators, P. Beardsley and S. Stefanovic for field assistance, and B. Watson for genotyping. We are grateful to F. Nicholson and the Carnegie Institution of Washington for allowing us to use the Mather field station. Y. Sam provided helpful comments on the manuscript. This work was supported by an award from the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.D.B. (toby@u.washington.edu).

Light-induced hormone conversion of T₄ to T₃ regulates photoperiodic response of gonads in birds

Takashi Yoshimura¹, Shinobu Yasuo¹, Miwa Watanabe¹, Masayuki Iigo³, Takashi Yamamura¹, Kanjun Hirunagi² & Shizufumi Ebihara¹

¹Division of Biomodeling, Graduate School of Bioagricultural Sciences, Nagoya University, and ²The Nagoya University Museum, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan

³Department of Applied Biological Chemistry, Faculty of Agriculture, Utsunomiya University, Mine-Machi, Utsunomiya, Tochigi 321-8505, Japan

Reproduction of many temperate zone birds is under photoperiodic control. The Japanese quail is an excellent model for studying the mechanism of photoperiodic time measurement because of its distinct and marked response to changing photoperiods. Studies on this animal have suggested that the mediobasal hypothalamus (MBH) is an important centre controlling photoperiodic time measurement^{1–8}. Here we report that expression in the MBH of the gene encoding type 2 iodothyronine deiodinase (Dio2), which catalyses the intracellular deiodination of thyroxine (T₄) prohormone to the active 3,5,3'-triiodothyronine (T₃), is induced by light in Japanese quail. Intracerebroventricular administration of T₃ mimics the photoperiodic response, whereas the Dio2 inhibitor iopanoic acid prevents gonadal growth. These findings demonstrate that light-induced Dio2 expression in the MBH may be involved in the photoperiodic response of gonads in Japanese quail.

The molecular mechanism of photoperiodic or seasonal time measurement is not well understood in any organism studied so far. In birds, the MBH—which includes the nucleus hypothalamicus posterior medialis (NHMP), the infundibular nucleus and the median eminence—is an important centre controlling photoperiodic time measurement (Supplementary Figs 1 and 2). For example, introduction of a lesion to the nucleus hypothalamicus posterior medialis and/or the infundibular nucleus resulted in loss of photoperiodic response of the gonads^{1–3} even though the gonadotrophin-releasing hormone (GnRH) system of the lesioned animal had been left intact⁴. Electrical stimulation of this area increases luteinizing hormone secretion⁵ and induces testicular growth⁶. Furthermore, c-Fos expression has been reported in these structures as a result of photostimulation for one long day (20/4 h light/dark cycle)^{7,8} and deep-brain photoreceptors are thought to be localized in the infundibular nucleus⁹. Recently, we have also observed the expression of circadian clock genes in the MBH, and proposed that the clock in the MBH may function as the 'photoperiodic clock'¹⁰. These observations indicate that all of the essential machinery for photoperiodic time measurement is localized in the MBH.

Single light pulses within the photo-inducible phase increase

serum luteinizing hormone concentration and cause photoperiodic response of the gonads¹¹. Therefore, it is expected that some molecular events take place in the MBH when animals are exposed to light within the photo-inducible phase. To identify genes that are responsible for the regulation of photoperiodic time measurement in birds, we performed a differential subtractive hybridization analysis. A total of 40 8-week-old male Japanese quail were raised under short day conditions (8/16 h light/dark cycle). From this group, 20 animals were exposed to a 1-h light pulse at zeitgeber time 14 (ZT14), whereas the other 20 animals were kept in darkness. One hour after the end of the light pulse (ZT16), both groups of animals were killed and the MBH was removed (Supplementary Fig. 1). Differential subtractive hybridization analysis was carried out and 150 clones were sequenced. Expression of all of these genes was verified using *in situ* hybridization. Among them, expression of only

one gene, *Dio2*, was significantly induced by the light pulse within the photo-inducible phase (ZT14) in the dorsal hypothalamus around the paraventricular organ (PVO) as well as the lateral hypothalamus¹² (Mann–Whitney *U*-test, $P < 0.01$; Fig. 1a–c). (More details of localization and the schematics of quail brain can be seen in Supplementary Fig. 2.) We then examined the effect of light exposure at times outside the photo-inducible phase (ZT9 and ZT21). However, light-induced expression of *Dio2* was not observed at ZT9 or ZT21 (Mann–Whitney *U*-test; $P > 0.05$), which indicates that the light-induced expression of *Dio2* is specific to the photo-inducible phase (Fig. 1c).

We next examined the expression profiles of *Dio2* in the dorsal hypothalamus under short and long day conditions. Unexpectedly, expression of *Dio2* in the dorsal hypothalamus was almost undetectable at any time of day, whether examined under long or short day conditions (Fig. 1d). Notably, expression of *Dio2* was observed in the infundibular nucleus and the median eminence (both of which are collectively known as the basal tuberal hypothalamus; BTH) (Fig. 1e, f; see also Supplementary Fig. 2). Consequently, we examined the effect of light on the expression of *Dio2* in these regions. Although acute induction of *Dio2* expression by light pulse was not detected at any of the phases examined (Mann–Whitney *U*-test, $P > 0.05$; Fig. 1g), significant differences in the level of expression between short day (weak expression) and long day (strong expression) conditions were observed in these regions (two-way analysis of variance (ANOVA), $F_{1,41} = 219.917$, $P < 0.0001$, * $P < 0.05$, Mann–Whitney *U*-test; Fig. 1e, f, h). No clear day/night difference in expression was observed in either long day or short day conditions (one-way ANOVA: long day, $F_{5,24} = 2.273$, $P = 0.0795$; short day, $F_{5,17} = 0.861$, $P = 0.5269$). It has been shown that introduction of a small lesion to the nucleus hypothalamicus posterior medialis (including the PVO) or the infundibular nucleus, or a combination of both lesions, completely blocks testicular development¹. The expression sites of *Dio2* revealed in our study are consistent with these results. Although it remains to be determined how these two structures are related to each other, these results suggest that expression of *Dio2* in both regions is of critical importance for the photoperiodic response of the gonads.

The enzyme *Dio2* converts T_4 to T_3 , which is primarily responsible for thyroid hormone action. *Dio2* has an essential role in the local control of levels of T_3 in the brain through mechanisms that operate under a variety of conditions to maintain T_3 concentrations within a narrow range¹³. To determine the levels of T_3 and T_4 in the

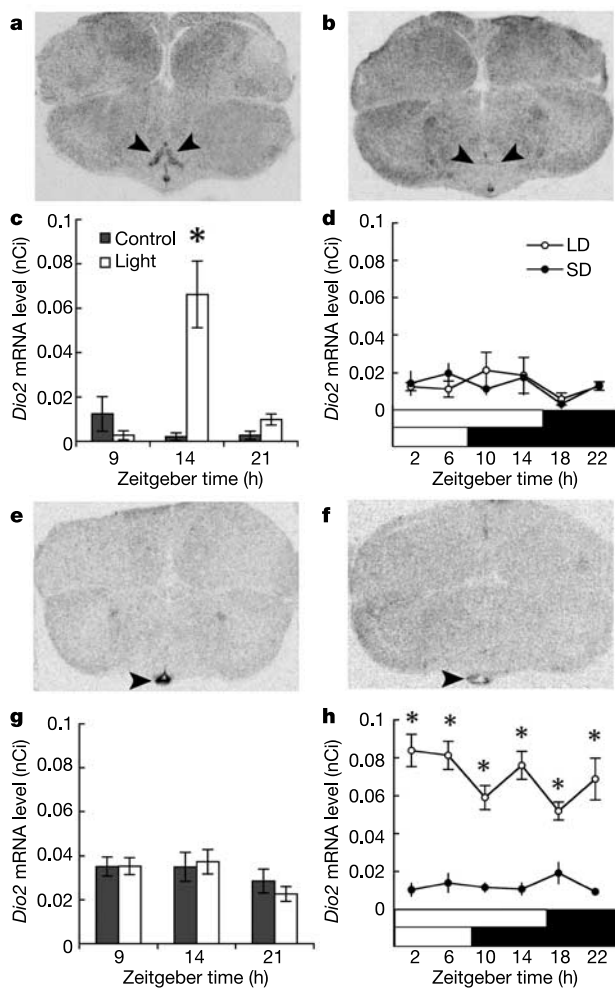


Figure 1 Identification of a gene that controls photoperiodic time measurement. **a, b**, Representative autoradiograms for light-induced expression of *Dio2* (**a**) and a control animal without light (**b**) in the dorsal hypothalamus (arrowhead). A single 1-h light pulse was given at ZT14, and brain was collected 1 h after the end of the light pulse. **c**, Light-induced *Dio2* expression in the dorsal hypothalamus was specific to the photo-inducible phase. Values are mean $\pm$ s.e.m. ($n = 5$). Asterisk, $P < 0.01$. **d**, Temporal expression profiles of the *Dio2* gene under short day (SD) and long day (LD) conditions in the dorsal hypothalamus. Values are mean $\pm$ s.e.m. ($n = 3$ –5). **e, f**, Representative autoradiograms of the BTH (arrowhead) of animals housed under long day (**e**) and short day (**f**) conditions. **g**, Effect of the light pulse on *Dio2* expression in the BTH. Values are mean $\pm$ s.e.m. ($n = 5$). **h**, Temporal expression profiles of the *Dio2* gene in the BTH in animals kept under short day and long day conditions. Values are mean $\pm$ s.e.m. ($n = 3$ –5). Asterisk, $P < 0.05$.

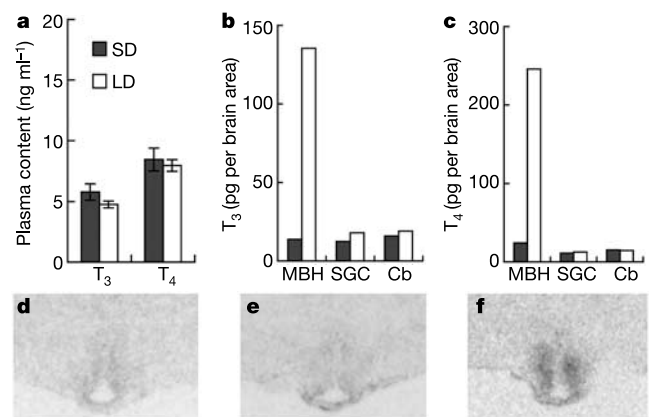


Figure 2 Locally generated T_3 acts on the BTH. **a**, Plasma content of T_3 and T_4 under long day and short day conditions. **b, c**, T_3 (**b**) and T_4 (**c**) contents are increased in the MBH under long day conditions, whereas they are not changed in the stratum griseum centrale (SGC) and the cerebellum (Cb). The MBH collected from ten animals were pooled. **d–f**, Expression of thyroid hormone receptor genes *Thra* (**d**), *ThrB* (**e**) and *Rxra* (**f**) in the BTH.

MBH, we measured the contents of pooled MBH collected from ten animals in each group under short day and long day conditions. Although the plasma concentration of T_3 and T_4 was not significantly different between short day and long day conditions (Mann–Whitney U -test, $P > 0.05$; Fig. 2a), the contents of both in the MBH were about tenfold higher in long day than in short day conditions (Fig. 2b, c). This difference was not observed in other parts of the brain, such as the stratum griseum centrale and the cerebellum (Fig. 2b, c). Recently it has been reported that expression of the gene encoding transthyretin (Ttr), which regulates thyroid hormone uptake, increases in Siberian hamster hypothalamus under long day conditions¹⁴. Although we could not detect expression of Ttr in the hypothalamus, strong expression was observed in the choroid plexus in Japanese quail; however, there was no difference in the expression of Ttr in the choroid plexus between long day and short day conditions (Supplementary Fig. 3). Therefore, the content of other type(s) of thyroid hormone transporters may increase in the MBH of Japanese quail under long day conditions that subsequently aid the generation of T_3 through increased T_4 uptake.

To explore the target site that locally generated T_3 acts upon, expression of thyroid hormone receptor genes ($Thra$, $Thrb$, $Thrb2$, $Rxra$ and $Rxrg$, which encode thyroid hormone receptor- α , - β , - $\beta 2$ and retinoid X receptor- α and - γ , respectively) were examined. Among these genes, weak expression of $Thra$ and $Thrb$ and strong expression of $Rxra$ was observed in the BTH, whereas expression of $Thrb2$ and $Rxrg$ was undetectable (Fig. 2d–f). Expression of $Thra$, $Thrb$ and $Rxra$ was observed for the entire day under both short day and long day conditions, and did not show a difference between day and night (data not shown). These results indicate that locally generated T_3 acts on the BTH. Although we have also examined expression of these genes in the preoptic area, where cell bodies of GnRH neurons are located, no signal was detectable under any conditions (data not shown). It has been suggested that photoperiodic GnRH release could be controlled at the GnRH terminals by glia^{8,15}, and thyroid hormones are known to have a critical involvement in the development, plasticity and function of the central nervous system¹³. Therefore, we examined the morphological changes in the median eminence under long day and short day conditions. Immunoelectron microscopy revealed that GnRH nerve terminals are in close proximity to the basal lamina in birds subjected to long day conditions (T.Ya., K.H., S.E. and T.Yo., unpublished observations). We also found that the GnRH terminal was enclosed by the end feet of glia in animals subjected to short day conditions (T.Ya., K.H., S.E. and T.Yo., unpublished observations). These morphological changes may allow the neurons to secrete GnRH owing to the increased access of their terminals to the

perivascular area surrounding the fenestrated portal capillaries under long day conditions.

To assess whether T_3 mediates the photoperiodic response of the gonads, we studied the effect of intracerebroventricular infusion of T_3 on testicular growth. Vehicle and several doses of T_3 and T_4 were infused into the third ventricle by an osmotic mini-pump, and the testicular size was measured before and after infusion. T_3 infusion induced testicular growth in a dose-dependent manner, even though the animals were kept under short day conditions (one-way ANOVA, $F_{5,18} = 4.008$, $P = 0.0128$; Fisher's LSD post-hoc test, $P < 0.05$), whereas T_4 infusion had only a minor effect (one-way ANOVA, $F_{3,16} = 0.32$, $P = 0.8111$; $n = 3$ –6; Fig. 3a). However, significant testicular growth was not observed in the largest dose of T_3 . Iopanoic acid is known to inhibit the conversion of T_4 to T_3 (ref. 16), thus we examined the effect of iopanoic acid infusion and found that it reduced testicular growth under long day conditions ($n = 4$ –5; Mann–Whitney U -test, $P < 0.05$; Fig. 3b).

Follett *et al.* have shown that peripheral injection of pharmacological doses of thyroid hormone (T_4 and T_3) can mimic photoperiodically induced gonadotrophin secretion and gonadal growth^{17,18}. In these previous studies, however, T_4 was more effective than T_3 in mimicking gonadotrophin release. The findings seem to stand in contradiction to our results. In blood, however, about 33% and 45% of T_4 is known to be converted to T_3 and reverse T_3 (rT_3), respectively¹⁹. Furthermore, T_3 is converted to 3,3'-diiodothyronine (3,3'- T_2) in the blood. Therefore, in the studies of Follett *et al.*, most of the peripherally delivered T_4 might be converted to T_3 and act on the central nervous system. In addition, we have shown that iopanoic acid, an inhibitor of Dio2, reduces testicular growth under long day conditions. Thus, our results are consistent with the hypothesis that Dio2 is important for regulation of the photoperiodic response of the gonads. Of note, T_3 infusion did not maximize testicular size and iopanoic acid did not block testicular growth completely. There are two possible explanations for this finding. One is that T_3 and iopanoic acid were subject to partial inhibition or partial loss through imprecise delivery. It is impossible to infuse at exactly the same location in all birds, and drugs may be diffused and metabolized through the cerebrospinal fluid in some. Variability in the effect of T_3 seems to support this idea. The other possibility is the existence of other regulatory pathways or compensatory mechanisms. This possibility is supported by previous studies showing that after thyroidectomy, quail can still respond to photostimulation when transferred from short day to long day conditions, although increases in their follicle-stimulating hormone levels and in the sizes of testes and the cloacal gland are attenuated^{20,21}. This is in contrast with the observation that after starlings and house sparrows are photoperiodically blind thyroidectomy^{22,23}. The exploration of this mechanism is important for our eventual understanding of the complete molecular machinery of photoperiodism. Although thyroid hormone is of critical importance in the regulation of photoperiodically induced gonadal growth, as shown here, this hormone is also known to be involved in the regulation of photorefractoriness (insensitivity to previously stimulatory day-length)^{15,24}. The observation that no significant testicular growth is induced at the highest dose of T_3 (Fig. 3) suggests the possibility that photorefractoriness occurs as a result of downregulation of thyroid hormone activity. This hypothesis remains to be tested.

Although exogenous thyroid hormones mimic the effect of long day conditions, their precise role remains uncertain¹⁵. The present study has determined the target site and molecular events of T_4 to T_3 conversion that take place in the brain. Finally, it is also of interest to note that multiple studies indicate that thyroid hormones are essential for the maintenance of seasonal reproductive changes in a number of mammals²⁵. Thus, our present study may provide the means for a collective understanding of the molecular mechanism of photoperiodic time measurement in all photoperiodic vertebrates. □

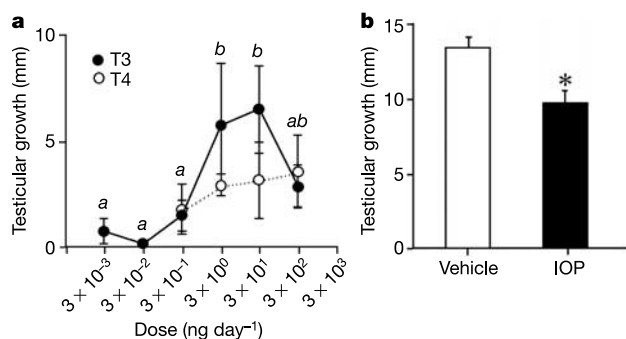


Figure 3 Intracerebroventricular T_3 infusion mimics photoperiodically induced testicular growth. **a**, T_3 and T_4 were infused under short day conditions. Testicular length was measured before and after infusion ($n = 3$ –6). Different characters (*a*, *b*) indicate significant differences ($P < 0.05$). **b**, Iopanoic acid infusion reduces testicular growth under long day conditions ($n = 4$ –5). Asterisk, $P < 0.05$.

Methods

Animals

Four-week-old male Japanese quail (*Coturnix coturnix japonica*) were obtained from a dealer and kept under short day conditions (8/16 h light/dark cycle) in light-tight boxes. The boxes were placed in a room at a temperature of $24 \pm 1^\circ\text{C}$. Light was supplied by fluorescent lamps with a light intensity of 200 lx measured at the head level of the birds. Short day animals were kept under short day conditions whereas long day animals were transferred from short to long day conditions (16/8 h light/dark) for 2 weeks. Food and water were available *ad libitum* and were replenished at least twice a week. Animals were treated in accordance with the guidelines of Nagoya University.

Differential subtractive hybridization analysis

Brain slices (3 mm) of quail were generated by mouse brain matrix (ASI), and the MBH was punched out (3 mm diameter). Total RNA was prepared from 20 pooled MBH using Trizol reagent (Gibco BRL). Poly(A)⁺ RNA was purified using oligotex-dt30 Super (Takara). Differential subtractive hybridization analysis was performed according to the manufacturer's instructions (PCR-select complementary DNA subtraction Kit, Clontech). Final PCR products were inserted into a TA cloning vector (Invitrogen) and sequenced by an ABIPrism 373 using the Big Dye Terminator kit (ABI).

In situ hybridization

Animals were killed by decapitation and the brains were immediately removed to avoid acute changes in gene expression. *In situ* hybridization was carried out as previously described²⁶. Antisense and sense 45-nucleotide oligonucleotide probes were labelled with [³³P]dATP (NEN) using terminal deoxyribonucleotidyl transferase (Gibco BRL). *Dio2*, 5'-GATGGTTCAGCCTCAATGAATATCAAGACGGAATACATTCTGTGA-3'; *Thra*, 5'-TTGATGGAATTGCGGTGAATGGAA CAGAAGCCAGCACCCTGGAC-3'; *Thrb*, 5'-CGGAGTGAGAGAACAGAAAATGAAGCTCTAGTAAGGTGGCAGTGG-3'; *Thrb2*, 5'-TGAAGTGCACCCAGCTGCTGTAGCAATTGCTACATGCAGTCCAC-3'; *Rxra*, 5'-GGCATGAGTTAAGACACGCGATGGACACCAACAACTTCCTGCCA-3'; *Rxrg*, 5'-ATTCCCTGTTTCATGCCAGCTCCAGCTCTGTAGCCCATCATCCA-3'; *Ttr*, 5'-ACAGTACTCGTTAGCTGCAGGACTTCCTCTGACTGCATCCAGCACT-3'.

Hybridization was carried out overnight at 42 °C. After glass slides were washed, they were air dried and apposed to Biomax-MR film (Eastman Kodak Co.) for 2 weeks with ¹⁴C standards (American Radiolabelled Chemicals). Relative optical densities were measured by using a computed image-analysing system (MCID Imaging Research), and were converted into the relative radioactive value (nCi) by ¹⁴C standards. Specific hybridization signals were obtained by subtracting background values obtained from adjacent brain areas that did not exhibit a hybridization signal.

Quantification of T₃ and T₄

Thyroid hormones in the brain were extracted with ethanol as described^{27,28}. Concentrations of T₄ and T₃ in the brain and plasma were determined by radioimmunoassay as described previously^{27,28}.

Intracerebroventricular infusion

Animals were raised under short day conditions to an age of 8 weeks. T₃ (T-2877, Sigma) and T₄ (T-2376, Sigma) were dissolved in a solution containing 0.001 M NaOH and 0.9% NaCl. Iopanoic acid (I 0300, Tokyo Kasei) was dissolved in 0.05 M NaOH and 0.2 M HCl was added. Intracerebroventricular infusion was carried out using an ALZET 2002 osmotic mini-pump with a brain infusion kit according to the manufacturer's instructions (ALZET). Animals infused with iopanoic acid were transferred into long day conditions 2 days after the beginning of the 2-week infusion period. Testicular size was measured before and 3 weeks after the beginning of infusion. Placement and patency of the canula were verified by injecting Evans blue dye after the experiment, as suggested by the manufacturer.

Received 14 July; accepted 25 September 2003; doi:10.1038/nature02117.

- Sharp, P. J. & Follett, B. K. The effect of hypothalamic lesions on gonadotrophin release in Japanese quail (*Coturnix coturnix japonica*). *Neuroendocrinology* **5**, 205–218 (1969).
- Davies, D. T. & Follett, B. K. The neuroendocrine control of gonadotrophin release in Japanese quail. I. The role of the tuberal hypothalamus. *Proc. R. Soc. Lond. B* **191**, 303–315 (1975).
- Ohta, M. & Homma, K. Detection of neural connections to the infundibular complex by partial or complete hypothalamic deafferentation in male quail. *Gen. Comp. Endocrinol.* **68**, 286–292 (1987).
- Juss, T. S. in *Avian Endocrinology* (ed. Sharp, P. J.) 47–60 (Soc. Endocrinol., Bristol, 1993).
- Konishi, H., Foster, R. G. & Follett, B. K. Evidence for a daily rhythmicity in the acute release of LH in response to electrical stimulation in the Japanese quail. *J. Comp. Physiol. A* **161**, 315–319 (1987).
- Ohta, M., Wada, M. & Homma, K. Induction of rapid testicular growth in quail by phasic electrical stimulation of the hypothalamic photosensitive area. *J. Comp. Physiol. A* **154**, 583–589 (1984).
- Meddle, S. L. & Follett, B. K. Photoperiodic activation of Fos-like immunoreactive protein in neurons within the tuberal hypothalamus of Japanese quail. *J. Comp. Physiol. A* **176**, 79–89 (1995).
- Meddle, S. L. & Follett, B. K. Photoperiodically driven changes in Fos expression within the basal tuberal hypothalamus and median eminence of Japanese quail. *J. Neurosci.* **17**, 8909–8918 (1997).
- Silver, R. *et al.* Coexpression of opsin- and VIP-like-immunoreactivity in CSF-contacting neurons of the avian brain. *Cell Tissue Res.* **253**, 189–198 (1988).
- Yasuo, S., Watanabe, M., Okabayashi, N., Ebihara, S. & Yoshimura, T. Circadian clock genes and photoperiodism: Comprehensive analysis of clock genes expression in the mediobasal hypothalamus, the suprachiasmatic nucleus and the pineal gland of Japanese quail under various light schedules. *Endocrinology* **144**, 3742–3748 (2003).
- Follett, B. K. & Sharp, P. J. Circadian rhythmicity in photoperiodically induced gonadotrophin release and gonadal growth in the quail. *Nature* **223**, 968–971 (1969).
- Kuenzel, W. J. & Masson, M. A. *Stereotaxic Atlas of the Brain of the Chick (Gallus domesticus)* (Johns Hopkins Univ. Press, Baltimore, 1988).

- Bernal, J. Action of thyroid hormone in brain. *J. Endocrinol. Invest.* **25**, 268–288 (2002).
- Prendergast, B. J., Mosinger, B. Jr, Kolattukudy, P. E. & Nelson, R. J. Hypothalamic gene expression in reproductively photorefractory and photorefractory Siberian hamsters. *Proc. Natl Acad. Sci. USA* **99**, 16291–16296 (2002).
- Dawson, A., King, V. M., Bentley, G. E. & Ball, G. F. Photoperiodic control of seasonality in birds. *J. Biol. Rhythms* **16**, 365–380 (2001).
- Leonard, J. L. & Visser, T. J. in *Thyroid Hormone Metabolism* (ed. Hennemann, G.) 189–229 (Marcel Dekker, New York, 1986).
- Follett, B. K. & Nicholls, T. J. Acute effect of thyroid hormones in mimicking photoperiodically induced release of gonadotropins in Japanese quail. *J. Comp. Physiol. B* **157**, 837–843 (1988).
- Follett, B. K., Nicholls, T. J. & Mayes, C. R. Thyroxine can mimic photoperiodically induced gonadal growth in Japanese quail. *J. Comp. Physiol. B* **157**, 829–835 (1988).
- Chopra, I. J. *et al.* Pathways of metabolism of thyroid hormones. *Recent Prog. Horm. Res.* **34**, 521–567 (1978).
- Follett, B. K. & Nicholls, T. J. Photorefractoriness in Japanese quail: possible involvement of the thyroid gland. *J. Exp. Zool.* **232**, 573–580 (1984).
- Follett, B. K. & Nicholls, T. J. Influences of thyroidectomy and thyroxine replacement on photoperiodically controlled reproduction in quail. *J. Endocrinol.* **107**, 211–221 (1985).
- Dawson, A. Thyroidectomy progressively renders the reproductive system of starlings (*Sturnus vulgaris*) unresponsive to changes in daylength. *J. Endocrinol.* **139**, 51–55 (1993).
- Dawson, A. Thyroidectomy of house sparrow (*Passer domesticus*) prevents photo-induced testicular growth but not the increased hypothalamic gonadotrophin-releasing hormone. *Gen. Comp. Endocrinol.* **110**, 196–200 (1998).
- Dawson, A. & Thapliyal, J. P. *Avian Endocrinology* (eds Dawson, A. & Chaturvedi, C. M.) 141–151 (Narosa, New Delhi, 2001).
- Nicholls, T. J., Follett, B. K., Goldsmith, A. R. & Pearson, H. Possible homologies between photorefractoriness in sheep and birds: the effect of thyroidectomy on the length of the ewe's breeding season. *Reprod. Nutr. Dev.* **28**, 375–385 (1988).
- Yoshimura, T. *et al.* Molecular analysis of avian circadian clock genes. *Mol. Brain Res.* **78**, 207–215 (2000).
- Tagawa, M. & Hirano, T. Presence of thyroxine in eggs and changes in its content during early development of chum salmon, *Oncorhynchus keta*. *Gen. Comp. Endocrinol.* **68**, 129–135 (1987).
- Ikuta, K., Aida, K., Okumoto, N. & Hanyu, I. Effects of sex steroids on the smoltification of masu salmon, *Oncorhynchus masou*. *Gen. Comp. Endocrinol.* **65**, 99–110 (1987).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank Nagoya University Radioisotope Center for use of its facilities. We also thank K. Aida, M. Tagawa and A. Munakata for providing antiserum, and A. Nishimura for technical assistance. This work was supported by the Program for Promotion of Basic Research Activities for Innovative Biosciences (PROBRAIN), a Grant-in-Aid for Encouragement of Young Scientists (to T.Y.), and a Grant-in-Aid for Scientific Research (to S.E.) from the Ministry of Education, Science, Sports and Culture.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.Yo. (takashiy@agr.nagoya-u.ac.jp).

APL regulates vascular tissue identity in *Arabidopsis*

Martin Bonke^{1*}, Siripong Thitamadee^{1*}, Ari Pekka Mähönen¹, Marie-Theres Hauser² & Ykä Helariutta¹

¹Plant Molecular Biology Laboratory, Institute of Biotechnology, POB 56, FIN-00014, University of Helsinki, Finland

²Center of Applied Genetics, BOKU - University of Natural Resources and Applied Life Sciences Vienna, Muthgasse 18, A-1190 Vienna, Austria

* These authors contributed equally to this work

Vascular plants have a long-distance transport system consisting of two tissue types with elongated cell files, phloem and xylem¹. Phloem has two basic cell types, enucleate sieve elements and companion cells. Xylem has various lignified cell types, such as tracheary elements, the differentiation of which involves deposition of elaborate cell wall thickenings and programmed cell death^{1–4}. Until now, little has been known about the genetic control of phloem–xylem patterning. Here we identify the *ALTERED PHLOEM DEVELOPMENT (APL)* gene, which encodes a MYB coiled-coil-type transcription factor that is required for

phloem identity in *Arabidopsis*. Phloem is established through asymmetric cell divisions and subsequent differentiation. We show that both processes are impaired by a recessive *apl* mutation. This is associated with the formation of cells that have xylem characteristics in the position of phloem. The *APL* expression profile is consistent with a key role in phloem development. Ectopic *APL* expression in the vascular bundle inhibits xylem development. Our studies suggest that *APL* has a dual role both in promoting phloem differentiation and in repressing xylem differentiation during vascular development.

The genetic control of vascular tissue identity is not well understood, although there is some genetic evidence for the involvement of a dorsiventral gradient specifying plant organs^{5,6}, as well as indirect physiological evidence for the involvement of various phytohormones (such as auxin, cytokinin and gibberellic acid (GA)) and/or sugars⁷. We report here the isolation and characterization of a novel *Arabidopsis* mutant defective in formation of the xylem–phloem pattern. A schematic of vascular tissue development is shown in Fig. 1a. The seedling lethal *apl* mutant displays a determinate root growth with only occasional branches, shoot

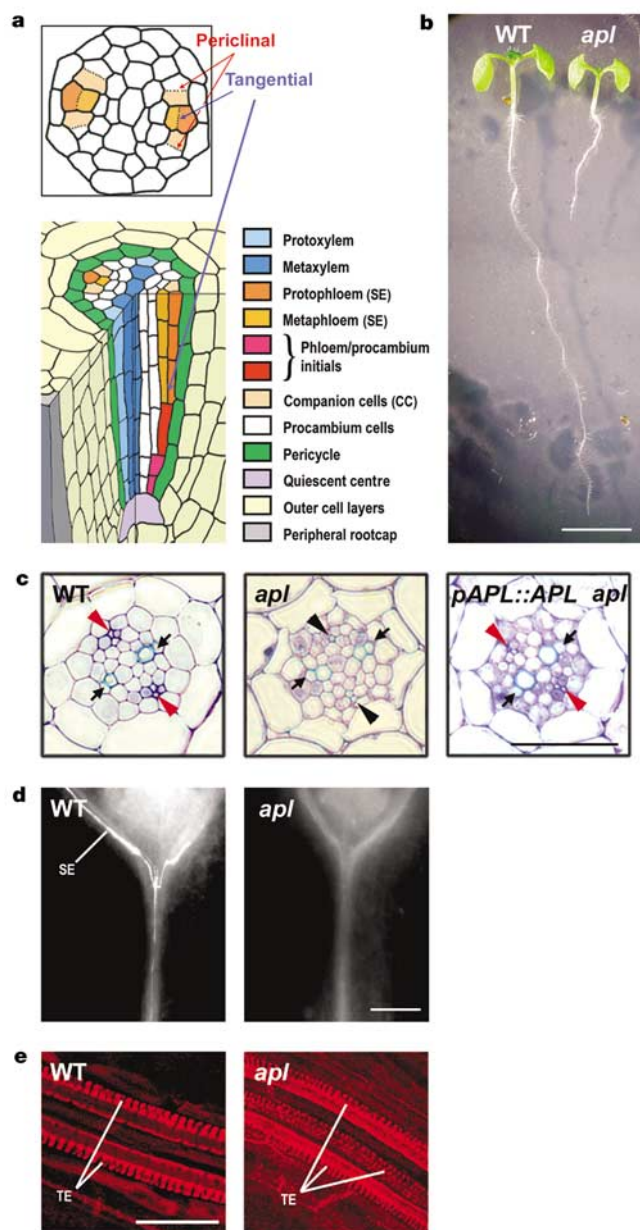


Figure 1 *apl* has an altered vascular pattern. **a**, Schematic of *Arabidopsis* root vasculature. The primary vascular pattern in *Arabidopsis* roots involves a xylem axis and two phloem poles. Phloem poles are formed by a tangential cell division (giving rise to sieve elements) and two periclinal divisions (resulting in companion cells). **b**, *apl* seedling. WT, wild type. **c**, Cross-sections of wild-type, *apl* and complemented *apl* primary roots, respectively. The position of protoxylem cells, SE and TE-like cells in the SE location in *apl* are indicated by arrows, red arrowheads and black arrowheads, respectively. **d**, *apl* seedlings stained with aniline blue show lack of callose deposition in shoot. **e**, Fuchsin staining indicates lignin deposition in an ectopic TE strand in *apl* root. Scale bars represent in **b**, 1 cm; **c**, **e**, 30 μ m; **d**, 100 μ m.

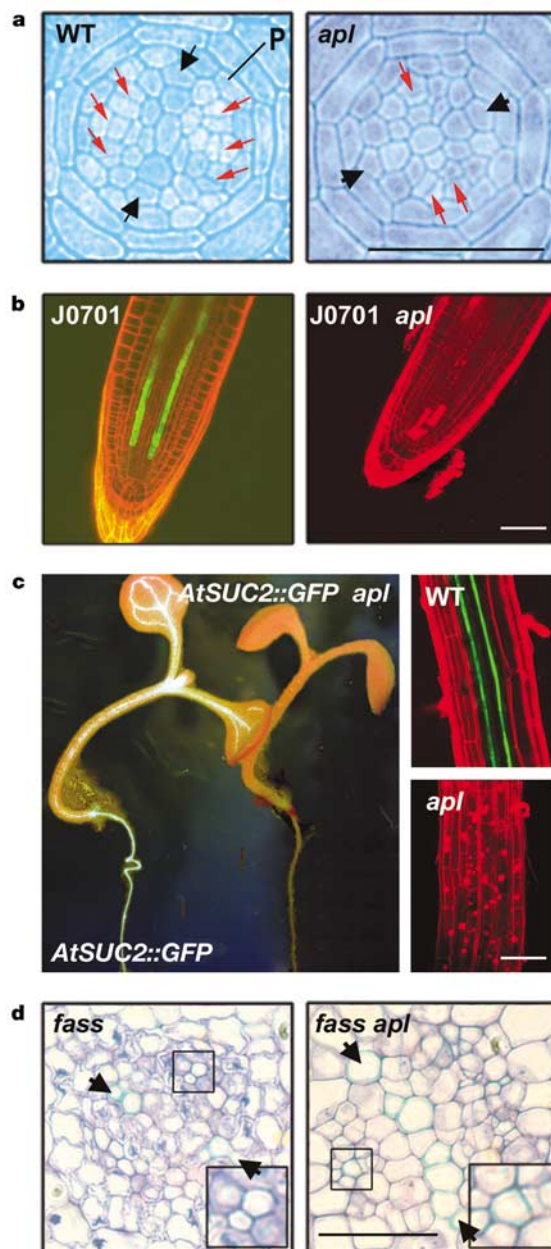


Figure 2 Asymmetric cell divisions, SE and CC differentiation are defective in *apl*. **a**, Root sections stained with toluidine blue at 27 μ m above the QC. Phloem-related divisions are indicated by red arrows. **b**, **c**, Protoxylem SE-specific GFP marker J0701 (Supplementary Fig. S2b) and CC-specific *AtSUC2pro::GFP* in wild type and *apl*. Signal is occasionally seen in patches of 1–3 cells, probably due to excision of the unstable *En-1* element (data not shown). **d**, Transverse sections of *fass apl* stained with toluidine blue show lignified cells in the prospective phloem pole. Black arrows indicate position of xylem axis. P, pericycle. Scale bars represent in **a**, 30 μ m; **b**, **c**, 60 μ m; **d**, 60 μ m.

development ultimately being arrested after formation of a few true leaves (Fig. 1b). Whereas the general pattern of the cell layers and other vascular cell files in an *apl* root is normal, neither sieve elements (SE) nor companion cells (CC) are evident in the location of the phloem poles (Fig. 1c). Furthermore, the characteristic callose deposition in SE of the cotyledon/hypocotyl junction⁸ is missing in *apl* seedlings (Fig. 1d). Strikingly, cells with characteristics of the tracheary elements (TE) of xylem are often observed in the positions normally occupied by SE (Fig. 1c, e). Phloem development is also specifically defective in the aerial organs (Fig. 1d; see leaf/hypocotyl in Supplementary Fig. S1a, b). There appears to be a slight delay (1 day in hypocotyl and 7 days in root) in the differentiation of the TE-like elements in the ectopic position compared with that in the normal position (which happens in both hypocotyl and root 1–2 days after germination).

To define the role of APL in phloem formation, we performed a detailed characterization of the *apl* mutant during root development. In wild type, phloem poles arise through a set of specific cell

divisions (Fig. 1a) with tangential divisions ($n = 12$; n indicates the number of division events analysed) giving rise to prospective SE cell files and periclinal divisions ($n = 24$) to the CC files^{9,10}. These divisions typically take place 50 μm above the quiescent centre (QC). In the *apl* vascular bundle, the timing of 3 out of 12 tangential and 12 out of 24 periclinal divisions seems to be delayed (Fig. 2a). In wild type, at about 70 μm above the QC, differentiation of the protophloem SE is evident. This differentiation process typically does not take place in *apl* (Supplementary Fig. S2a). In addition, the GFP signal specific for the protophloem SE after the tangential division in line J0701 is not detected in the *apl* background (Fig. 2b). Furthermore, expression of the CC-specific *AtSUC2pro::GFP*¹¹ is not observed in any part of the mutant line (Fig. 2c). In contrast to the phloem determinants, the various markers for xylem and xylem-associated cells appear as in wild type (Supplementary Fig. S2c, d). To investigate whether the defects in cell differentiation of *apl* seedlings are dependent on problems in the phloem-related divisions, we introduced *fass* into the *apl* background. The *fass* mutant

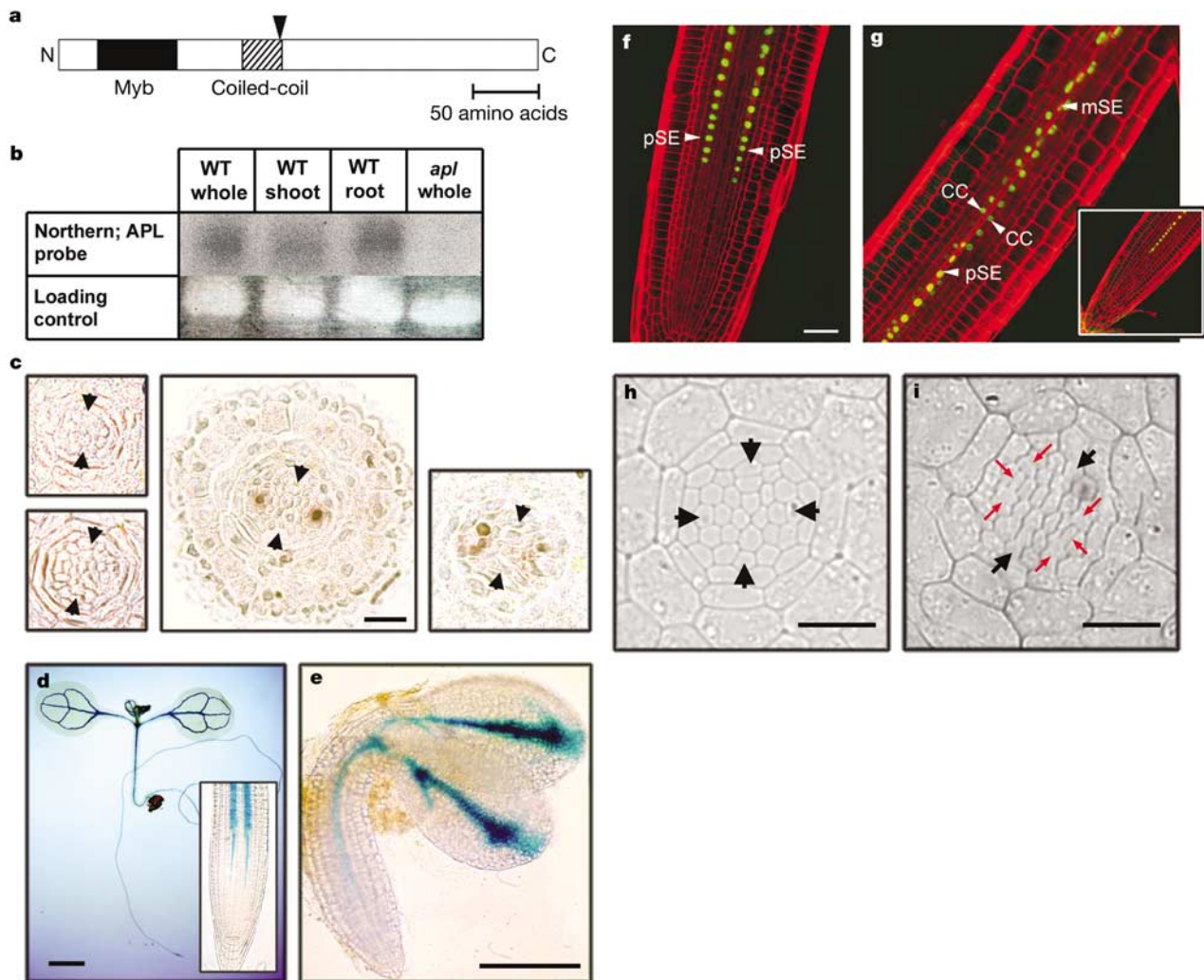


Figure 3 Molecular characterization of the *APL* locus. **a**, Structure of the putative *APL* transcription factor. The *En-1* transposon insertion site in *apl* is indicated by a black arrowhead. **b**, Northern hybridization analysis of *APL* expression. 18S rRNA is the loading control. **c**, *In situ* hybridization in serial sections. *APL* expression is not detected below the asymmetric cell divisions (top left), nor right above them (bottom left), but only slightly above the asymmetric cell divisions, first in the developing protophloem SE (middle) and higher up in the developing CC and metaphloem SE (right). A sense probe showed nonspecific staining (data not shown). The position of the xylem axis is indicated by the arrowheads. **d**, **e**, *APL:pro::GUS* expression in seedling (**d**) and during the bent-cotyledon

stage of embryogenesis (**e**). Insert in **d**, proximal to the root meristem. **f**, **g**, GFP-*APL* in developing SE cell files (**f** and **g**) and CC (**g**). Perpendicular to the observation plane in **f**, loss of GFP localization in mature protophloem SE (pSE) is accompanied by strong GFP expression in neighbouring CC (**g**). In addition, reappearing GFP signal in SE cell file at the top of the image indicates immature metaphloem SE (mSE) localization. Inset displays the tip beneath (**g**). **h**, **i**, Cross-section of nearly mature embryonic root (**h**) and the upper hypocotyl region (**i**). Black arrows indicate axis. Red arrows indicate the pattern conferred by the asymmetric cell divisions of phloem. Scale bars represent in **c**, **h**, **i**, 20 μm ; **d**, 0.4 cm; **e**, 60 μm ; **f**, **g**, 40 μm .

undergoes excessive numbers of cell divisions in all tissues¹² and is known to rescue the *wooden-leg* (*wol*) mutant, in which reduced cell proliferation during vascular development is accompanied by exclusive protoxylem differentiation. This indicates that *wol* affects morphogenesis rather than pattern formation¹³. The *fass apl* double mutant phenotype is additive; more vascular cells are present but TE-like cells are still found in the position of phloem (Fig. 2d). We conclude that APL is required for the asymmetric cell divisions as well as for SE and CC differentiation during phloem development, and that the absence of APL function results in differentiation of TE-like cells at the site of the phloem pole.

To identify the *APL* locus at a molecular level, we determined that the *apl* mutation resulted from an *En-1* transposition event that disrupts the fifth exon of the gene *At1g79430* (supported by a full-length complementary DNA, accession number NM_179574). The putative APL protein has 358 amino acid residues and contains a MYB-related DNA-binding domain and a predicted coiled-coil dimerization domain¹⁴ (Fig. 3a). A 6,535-base-pair (bp) genomic DNA fragment containing only the *APL* gene can rescue the *apl* mutant phenotype (Fig. 1c). Cosegregation of the *En-1* transposon with the *apl* mutation and an excision event of *En-1* confirmed the localization (see Methods). The insertion of *En-1* results in a truncated form of the putative APL protein (159 amino-acid residues), containing MYB and only part of the coiled-coil domain. These data, and the northern hybridization analysis showing that *APL* messenger RNA does not accumulate in the *apl* background (Fig. 3b), indicate that the *apl* mutation results in loss of function.

Next, we examined the developmental expression pattern of the *APL* gene. Based on northern analysis, *APL* is expressed in both shoot and root (Fig. 3b). To localize the expression at a cellular level during root development, we performed *in situ* hybridization. We found that *APL* is first specifically expressed in the developing protophloem SE soon after the phloem-specific cell divisions have taken place, then it is expressed slightly higher up in the CC and metaphloem SE (Fig. 3c). Phloem-specific expression of *APLpromoter::GUS* was observed throughout the vascular strands in *Arabidopsis* seedlings (Fig. 3d). Consistent with the *APL* expression pattern close to the meristem detected by *in situ* hybridization, *APLpro::GUS* expression in the primary root is first restricted to the cell files of protophloem SE (Fig. 3d; Supplementary Fig. S3a). A pattern involving a broader expression domain of *GUS* was observed 7–8 cells higher.

We next investigated the subcellular localization of APL by constructing a fusion of the green fluorescent protein (*GFP*) gene and *APL*, placed under the regulation of the *APL* promoter. The *APLpro::GFP-APL* fusion construct was able to rescue the developmentally arrested phenotype of *apl*, indicating that the fusion protein retained APL activity (data not shown). *GFP-APL* fluorescence was observed in the nucleus of the developing SE and CC in a dynamic pattern (Fig. 3f, g; Supplementary Fig. S3b), mirroring the temporal and spatial differentiation events of the two phloem cell types.

To investigate when phloem identity is established, we analysed the status of vascular anatomy and *APL* expression during embryogenesis. *APLpro::GUS* was detected in the prospective phloem tissue of embryonic cotyledons and hypocotyl in a nearly mature, but not a torpedo-stage, embryo (Fig. 3e; data not shown). Before *APL* expression, a distinct vascular pattern consisting of four nearly identical poles was observed in hypocotyl and in root (Fig. 3h; Supplementary Fig. S3c). In embryonic hypocotyl, after the onset of *APL* expression a somewhat variable anatomical pattern only occasionally similar to that of the postembryonic root was observed (Fig. 3i; compare with Fig. 2a). Thus, compared with the patterning of the outer layers (which has occurred by the torpedo stage^{13,15–18}), the patterning of phloem starts relatively late during embryogenesis and is completed only postembryonically. Although the developmental expression pattern of *APL* suggests that APL may regulate phloem identity throughout plant development, the highly variable

nature of the embryonic pattern in wild type precludes a conclusive phenotypic analysis of the *apl* embryos.

Based on the presence of TE-like cells in the *apl* phloem poles, it appears that APL may be capable of inhibiting the differentiation of TE-like identity characteristic of xylem. To investigate this, we expressed *APL* throughout the procambial region of the root under regulation of the *CRE1/WOL/AHK4* promoter, *WOLpro*^{9,19,20} (Supplementary Fig. S4a–c). *WOLpro::APL* transgenic seedlings exhibited normal growth and cellular patterns (Supplementary Fig. S4d). However, the lignin deposition pattern and cell viability revealed a striking phenotype during xylem development (Fig. 4a, b; Supplementary Fig. S4e). Mirroring the strong activity of the *WOL* promoter near the root tip, cells that normally differentiate as TE in the protoxylem position remain undifferentiated in the *WOLpro::APL* plants (Fig. 4c, d). Furthermore, higher up in the

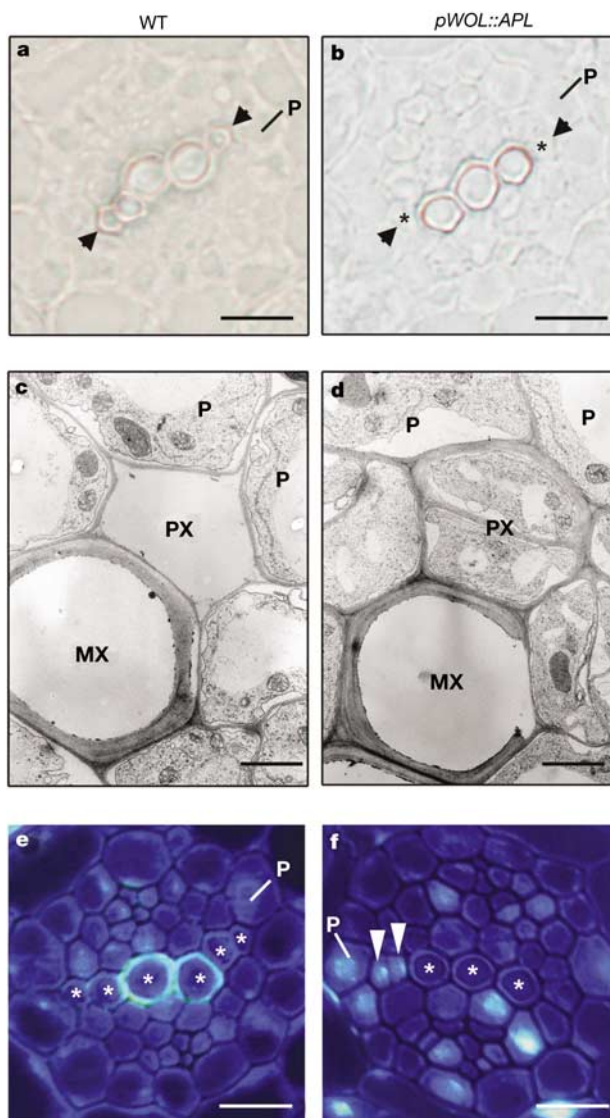


Figure 4 Ectopic *APL* expression inhibits xylem differentiation. **a, b**, Phloroglucinol-stained cross-sections of roots indicate that lignification of the cells at protoxylem position does not occur in *WOLpro::APL* roots. Xylem axis indicated by arrowheads. Protoxylem position indicated by asterisk. **c, d**, Cell in protoxylem position does not undergo autolysis in *WOLpro::APL* plants (and sometimes divides tangentially). **e, f**, The overlaid brightfield and ultraviolet fluorescence images of DAPI (4,6-diamino-2-phenylindole) staining show the presence of nuclei in the cell at the protoxylem position (arrowheads). TE-like cells are indicated by asterisks. P, pericycle; PX, protoxylem position; MX, metaxylem position. Scale bars represent in **a, b, e, f**, 20 μ m; **c, d**, 500 nm.

centre of the vascular bundle, cells in the metaxylem position differentiate later than those in the wild type (Supplementary Fig. S4e). It is unclear why cells in the protoxylem position remain specifically undifferentiated even further up. We conclude that *WOLpro::APL* is sufficient to inhibit differentiation of TE. Consistent with this model of the repressive activity of APL on an early step of xylem differentiation, *WOLpro::APL* can suppress the all-protoxylem vascular pattern of *wol* roots, resulting in an array of phloem–xylem patterns (data not shown). On the basis of anatomical analysis and detection of a nucleus in the undifferentiated cells, it appears that APL itself is not sufficient to promote phloem development in an ectopic location (Fig. 4e, f; Supplementary Fig. S4f–i).

We believe that APL is the first gene function to be identified that is specifically required for vascular identity in plants. The mutant analysis indicates a role for APL during both phloem-related cell divisions and cell differentiation. However, the spatial expression pattern specific for the developing SE and CC somewhat higher up than the dividing cells indicates that APL may be required primarily for SE and CC differentiation. In this scenario, the defects in phloem-related cell divisions may be secondary, either because of the absence of an as yet unknown cell non-autonomous action of APL, or alternatively because of a metabolic problem associated with abnormal phloem differentiation in *apl* cells. The contrasting nature of the *apl* phenotype (with TE-like cells in positions normally occupied by phloem) and the *WOLpro::APL* phenotype (with undifferentiated cells in the protoxylem position) suggests that in addition to promoting phloem differentiation, APL may also be required for inhibiting xylem differentiation in phloem poles during vascular development. An analogous concept has been described for Notch5 during artery–vein pattern formation in zebrafish; it is required for arterial development, and when ectopically expressed, can repress venous cell fate^{21,22}. The targets (some of them possibly related to phytohormone-mediated functions^{7,23,24}) of the positive and repressive functions of APL remain to be identified. Furthermore, it remains to be determined whether vascular identity is influenced through APL by the dorsoventral regulation of organ polarity in plants. □

Methods

Molecular characterization of APL

The *apl* mutant (selected from an *En-1* mutagenized pool²⁵) was inherited as a recessive mendelian trait (507_{wc}:185_{apl}). For mapping, *apl* (Columbia ecotype) was crossed to Ler plant. Analysis of 1,020 F₂ plants delimited APL to a 98-kilobase (kb) region between the newly generated markers at position YUP8H12R-78,178 kb and T8K14-56,936 kb. In addition, two TAC clones²⁶ (88N24 and 49O23) representing a 30-kb region in this window complemented the mutant phenotype.

To locate the *En-1*-tagged position, we designed specific oligonucleotide primer pairs to amplify by polymerase chain reaction (PCR) the overlapping fragments that spanned this window. One of the primer pairs showed the expected PCR product (1,576 bp) in Colombia ecotype, whereas the amplification of genomic DNA in *apl* was undetectable, suggesting that the transposon insertion was in this region. To identify the *En-1*-tagged position, the genomic PCR product from *apl* (851 bp) using the *En-1*-specific primer 5′-TCAGGCTCACATCATGCTAGTCC-3′ and the gene-specific primer 5′-TTGTTTGA TCAGGATGCAATGG-3′ was sequenced.

Cosegregation analysis was performed using the primers indicated above, and the genotype of the individual plants was determined by the segregation ratio of their progeny. The resultant PCR fragment cosegregated with *apl* plants (*n* = 10) and *APL/apl* heterozygous plants (*n* = 8), whereas no detectable PCR product containing the *En-1* transposon fragment was observed in wild-type plants (*n* = 6). Furthermore, an *En-1* excision event restoring the open reading frame of the APL gene was analysed. For complementation analysis, the 6,535-bp *PvuII*–*Bam*HI genomic DNA fragment was cloned into the pRD400 binary vector and introduced into *Agrobacterium* strain C58C1 GV3101. The construct was transformed into heterozygous *APL/apl* plants. Independent T2 plants were genotyped and were analysed for the segregation ratio using PCR.

To generate the *WOLpro::APL* construct, a *Hind*III genomic fragment of APL containing the entire coding region was subcloned from TAC 88N24 into pBluescript II. A *Sma*I site was introduced immediately before the ATG start codon, and a *Sac*I site was introduced at position 144 after the TGA stop codon. The resulting *Sma*I–*Sac*I fragment containing the 1.76-kb APL coding region was fused to the 3′ fragment of the 2.7-kb *WOL* promoter region in binary vector pBIN19. The plasmids were electroporated into *Agrobacterium* strain GV3101, which was used to transform an *Arabidopsis* plant by the floral dip method. Transgenic lines were selected on MS media supplemented with 50 µg ml^{−1} kanamycin. Multiple transgenic lines were obtained, and two independent

transgenic lines in which the transgene inserted at a single locus were selected for further characterization. The status of APL expression in these seedlings was confirmed by *in situ* hybridization (Supplementary Fig. S4c).

To construct *APLpro::GUS*, the 2.9-kb APL 5′ upstream region was cloned as a *Sall*–*Sma*I fragment into pBI101. To make the in-frame fusion of APL and GFP, we introduced restriction sites at both ends of coding sequences for APL and GFP. A linker sequence encoding triple copies of glycine was inserted between APL and GFP. The β-glucuronidase (GUS) fragment from an *APLpro::GUS* construct was replaced by the GFP–APL fragment.

The GFP marker lines used in this study are described at <http://www.plantsci.cam.ac.uk/Haseloff/CATALOGUES/Jlines/index.html> and were obtained through the Nottingham Arabidopsis Stock Centre.

Histology and northern and *in situ* hybridization

Plastic embedding, sectioning, staining, photography, fuchsin staining and confocal imaging were performed as described⁹. The anatomical studies on phloem cell division and differentiation were carried out on 3-day-old seedlings. For the anatomical analysis of ectopic xylem formation and complementation, 9-day-old seedlings were used. For northern and *in situ* hybridization experiments, 7-day-old seedlings were prepared and sectioned as described⁹. Seedlings were grown on Petri plates containing 0.5X Murashige and Skoog salt mixture and 1% sucrose, pH 5.8, in 0.8% agar. APL gene-specific hydrolysed sense- and antisense probes representing a non-conserved region in the 3′ end of the APL gene (757 bp, 5′-CTTCTGCCATGGATATTCAGCG-3′, 5′-CACCCAAATGG CGAGTTTCTTCC-3′) were labelled using the DIG RNA labelling kit (Boehringer) according to the manufacturer's protocol. Images were processed with Adobe Photoshop 6.0.1.

Received 17 July; accepted 29 September 2003; doi:10.1038/nature02100.

- Esau, K. *Anatomy of Seed Plants* 2nd edn (John Wiley & Sons, New York, 1977).
- Oparka, K. J. & Turgeon, R. Sieve elements and companion cells—traffic control centers of the phloem. *Plant Cell* **11**, 739–750 (1999).
- Kuhn, C. *et al.* Macromolecular trafficking indicated by localization and turnover of sucrose transporters in enucleate sieve elements. *Science* **275**, 1298–1300 (1997).
- Ye, Z. H. Vascular tissue differentiation and pattern formation in plants. *Annu. Rev. Plant Biol.* **53**, 183–202 (2002).
- McConnell, J. R. & Barton, M. K. Leaf polarity and meristem formation in *Arabidopsis*. *Development* **125**, 2935–2942 (1998).
- Waites, R. & Hudson, A. *phantastica*: a gene required for dorsoventrality of leaves in *Antirrhinum majus*. *Development* **121**, 2143–2154 (1995).
- Aloni, R. Differentiation of vascular tissues. *Annu. Rev. Plant Physiol.* **38**, 179–204 (1987).
- Carland, F. M. *et al.* Genetic regulation of vascular tissue patterning in *Arabidopsis*. *Plant Cell* **11**, 2123–2137 (1999).
- Mahonen, A. P. *et al.* A novel two-component hybrid molecule regulates vascular morphogenesis of the *Arabidopsis* root. *Genes Dev.* **14**, 2938–2943 (2000).
- Baum, S. F., Dubrovsky, J. G. & Rost, T. L. Apical organization and maturation of the cortex and vascular cylinder in *Arabidopsis thaliana* (Brassicaceae) roots. *Am. J. Bot.* **89**, 908–920 (2002).
- Imlau, A., Truernit, E. & Sauer, N. Cell-to-cell and long-distance trafficking of the green fluorescent protein in the phloem and symplastic unloading of the protein into sink tissues. *Plant Cell* **11**, 309–322 (1999).
- Torres-Ruiz, R. A. & Jurgens, G. Mutations in the *FASS* gene uncouple pattern formation and morphogenesis in *Arabidopsis* development. *Development* **120**, 2967–2978 (1994).
- Scheres, B. *et al.* Mutations affecting the radial organization of the *Arabidopsis* root display specific defects throughout the embryonic axis. *Development* **121**, 53–62 (1995).
- Rubio, V. *et al.* A conserved MYB transcription factor involved in phosphate starvation signaling both in vascular plants and in unicellular algae. *Genes Dev.* **15**, 2122–2133 (2001).
- Scheres, B. *et al.* Embryonic origin of the *Arabidopsis* primary root and root meristem initials. *Development* **120**, 2475–2487 (1994).
- Di Laurenzio, L. *et al.* The *SCARECROW* gene regulates an asymmetric cell division that is essential for generating the radial organization of the *Arabidopsis* root. *Cell* **86**, 423–433 (1996).
- Helariutta, Y. *et al.* *SHORT-ROOT* gene controls radial patterning of the *Arabidopsis* root through radial signaling. *Cell* **101**, 555–567 (2000).
- Lin, Y. & Schiefelbein, J. Embryonic control of epidermal cell patterning in the root and hypocotyl of *Arabidopsis*. *Development* **128**, 3697–3705 (2001).
- Inoue, T. *et al.* Identification of CRE1 as a cytokinin receptor from *Arabidopsis*. *Nature* **409**, 1060–1063 (2001).
- Suzuki, T. *et al.* The *Arabidopsis* sensor His-kinase, AHK4, can respond to cytokinins. *Plant Cell Physiol.* **42**, 107–113 (2001).
- Lawson, N. D. *et al.* Notch signaling is required for arterial–venous differentiation during embryonic vascular development. *Development* **128**, 3675–3683 (2001).
- Lawson, N. D. & Weinstein, B. M. Arteries and veins: making a difference with zebrafish. *Nature Rev. Genet.* **3**, 674–682 (2002).
- Yoshimura, T., Demura, T., Igarashi, M. & Fukuda, H. Differential expression of three genes for different β-tubulin isoforms during the initial culture of *Zinnia* mesophyll cells that divide and differentiate into tracheary elements. *Plant Cell Physiol.* **37**, 1167–1176 (1996).
- McCann, M. Tracheary element formation: building up to a dead end. *Trends Plant Sci.* **2**, 333–338 (1997).
- Wisman, E., Cardon, G. H., Fransz, P. & Saedler, H. The behaviour of the autonomous maize transposable element *En/Spm* in *Arabidopsis thaliana* allows efficient mutagenesis. *Plant Mol. Biol.* **37**, 989–999 (1998).
- Liu, Y. G. *et al.* Complementation of plant mutants with large genomic DNA fragments by a transformation-competent artificial chromosome vector accelerates positional cloning. *Proc. Natl Acad. Sci. USA* **96**, 6535–6540 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Kivimäki, K. Kainulainen, A. Hakonen, J. Immanen, Q. Yu, F. D. Rodriguez, G. Resch and M. Iljin-Jug for technical assistance and plant cultivation; E. Jokitalo for TEM; M. Laxell and S. Tähtiharju for material; I. Bancroft and K. van de Sande at GARNet/GeTCID; N. Sauer for the AtSUC2::GFP line; E. Mellerowicz, F. Rook and M. Saarma for advice; and P. N. Benfey, B. Scheres and C. Alonso Blanco for comments on the manuscript. Financial support was provided by the Academy of Finland, Tekes, EU, EMBO and the University of Helsinki.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Y.H. (yheliariu@operon.helsinki.fi).

An NS3 protease inhibitor with antiviral effects in humans infected with hepatitis C virus

Daniel Lamarre¹, Paul C. Anderson², Murray Bailey², Pierre Beaulieu², Gordon Bolger¹, Pierre Bonneau¹, Michael Bös², Dale R. Cameron^{2*}, Mireille Cartier¹, Michael G. Cordingley¹, Anne-Marie Faucher², Nathalie Goudreau², Stephen H. Kawai², George Kukolj¹, Lisette Lagacé¹, Steven R. LaPlante², Hans Narjes², Marc-André Poupart², Jean Rancourt², Roel E. Sentjens⁴, Roger St George⁵, Bruno Simoneau², Gerhard Steinmann³, Diane Thibeault¹, Youla S. Tsantrizos², Steven M. Weldon⁵, Chan-Loi Yong⁵ & Montse Llinàs-Brunet²

¹Departments of Biological Sciences and ²Chemistry, Research and Development, Boehringer Ingelheim (Canada) Ltd, Laval, Québec, H7S 2G5, Canada

³Clinical Research, Boehringer Ingelheim Pharma KG, Biberach 88397, Germany

⁴Academisch Medisch Center, 1105 AZ, Amsterdam, The Netherlands

⁵Research and Development, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, Connecticut 06877-0368, USA

* Present address: Micrologix Biotech Inc., BC Research Complex, 3650 Wesbrook Mall, Vancouver, British Columbia, V6S 2L2, Canada

Hepatitis C virus (HCV) infection is a serious cause of chronic liver disease worldwide with more than 170 million infected individuals at risk of developing significant morbidity and mortality^{1–3}. Current interferon-based therapies⁴ are suboptimal especially in patients infected with HCV genotype 1, and they are poorly tolerated, highlighting the unmet medical need for new therapeutics^{5,6}. The HCV-encoded NS3 protease is essential for viral replication^{7,8} and has long been considered an attractive target for therapeutic intervention in HCV-infected patients. Here we identify a class of specific and potent NS3 protease inhibitors and report the evaluation of BILN 2061, a small molecule inhibitor biologically available through oral ingestion and the first of its class in human trials. Administration of BILN 2061 to patients infected with HCV genotype 1 for 2 days resulted in an impressive reduction of HCV RNA plasma levels, and established proof-of-concept in humans for an HCV NS3 protease inhibitor. Our results further illustrate the potential of the viral-enzyme-targeted drug discovery approach for the development of new HCV therapeutics.

In spite of intensive efforts, many aspects of the disease and the biology of HCV remain unclear, making drug discovery an extremely challenging endeavour. Robust and practical cell culture systems that support viral replication are still lacking, necessitating the use of surrogate cellular systems such as the subgenomic HCV replicon cell model⁹ in support of drug discovery. The HCV-infected chimpanzee, an impractical model for routine studies of HCV biology, was until recently¹⁰ the only animal model supporting HCV replication. Target-based antiviral drug discovery—an

approach exemplified by the human immunodeficiency virus (HIV) drug discovery model, which mainly relies on the use of *in vitro* assays—has led recently to the identification of several anti-HCV compounds awaiting clinical validation through tangible therapeutic benefit in HCV-infected patients.

A substrate-based approach to the design of active site inhibitors of viral proteases generally benefits from a genetic conservation of the substrate-binding site and gains from available structural information defining the topographies of the complementary surfaces between the substrate or inhibitor and the protease. Substrate recognition by the NS3 protease is characterized by preference for a cysteine residue at P1 (nomenclature as in ref. 11) but requires extended interactions between enzyme and substrate distributed over the minimal substrate peptide P6 to P4'. This requirement is dictated by the shallow, solvent-exposed substrate-binding cleft of the enzyme. The NS3 protease therefore represented a formidable challenge for the design of small molecule inhibitors. Amino-terminal products derived from cleavage of peptide substrates are competitive inhibitors^{12,13} and the weak hexapeptide inhibitor Asp-Asp-Ile-Val-Pro-Cys (K_i of 79 μ M) served as the starting lead molecule for design efforts. Our initial lead optimization capitalized on maintaining the carboxy-terminal carboxylate functionality. This carboxylate confers increased potency against NS3 protease and selectivity with respect to a large panel of proteases¹⁴. Knowing the important contribution of the P1 residue in substrate and inhibitor binding to the enzyme, a large effort was devoted to the replacement of the reactive cysteine side chain. Evaluation of different alkyl and cycloalkyl side chains¹⁵ resulted in the identification of the (1R,2S)-1-amino-2-vinylcyclopropyl carboxylic acid as a suitable and chemically stable cysteine replacement (J.R. *et al.*, unpublished observation). Exploration of the P2 position identified key substituted proline derivatives^{16,17} that, in combination with optimal groups at P3 to P6, led to very specific and potent hexapeptide inhibitors with activity in the subgenomic HCV replicon model¹⁸. After an iterative and extensive structure–activity–relationship approach and guidance from NMR- and X-ray-derived structural information^{17,19,20}, tripeptide mimetics were identified that maintained potent and specific activity against the NS3 protease²¹. Rigidification of their scaffold by intramolecular linking of the P1 side chain to the P3 side chain produced novel macrocyclic inhibitors with desirable drug-like properties²² that exhibited improved NS3 protease inhibition in cells. Subsequently, we undertook an intensive structure–activity–relationship campaign leading to a subset of compounds that achieved the targeted

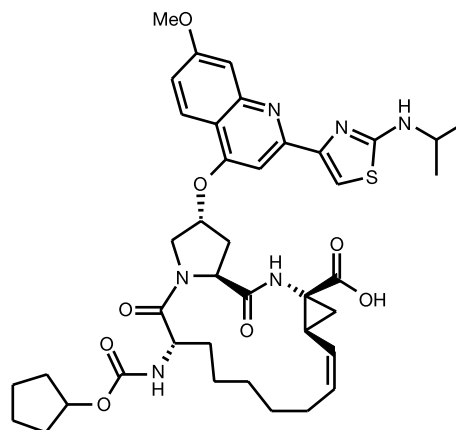


Figure 1 Chemical structure of BILN 2061. The synthesis of the macrocyclic inhibitor BILN 2061 has been described³⁰ and further details will be published elsewhere.

low-nanomolar cellular potency and an adequate oral pharmacokinetic profile in animals. BILN 2061 (Fig. 1) was selected from the optimized macrocyclic inhibitor series for further development.

BILN 2061 displayed potent and competitive inhibition of the NS3 proteases of HCV genotypes 1a and 1b with a mean K_i of 0.30 nM and 0.66 nM, respectively (Table 1). The inhibition of NS3 protease of HCV genotype 1b by BILN 2061 was reversible as demonstrated by the increase in steady-state velocity after dilution of a preformed inhibitor–enzyme complex into buffer-containing substrate (Supplementary Information 1). The inhibition of BILN 2061 was highly specific to the NS3 protease as demonstrated by the lack of significant activity (half-maximal inhibitory concentration (IC_{50}) > 30 μ M) against human leukocyte elastase and human liver cathepsin B, representatives of serine and cysteine proteases respectively. The ability of BILN 2061 to inhibit NS3 protease activity in human liver cells was evaluated using the subgenomic HCV replicon cell model (Table 1). Treatment of replicon-containing cells with BILN 2061 for 3 days resulted in a dose-dependent decrease of HCV RNA of two orders of magnitude with a mean 50% effective concentration (EC_{50}) of 4 nM and 3 nM for the HCV replicon 1a and 1b, respectively. The addition of 50% human serum to the culture medium resulted in less than a tenfold increase in EC_{50} . A mean 50% cytotoxic concentration (CC_{50}) of 33 μ M was observed for BILN 2061, resulting in an apparent selectivity index of 10,000 in Huh-7 cells when compared with the EC_{50} value obtained for inhibition of subgenomic HCV RNA replication. The mechanism of inhibition of BILN 2061 was further confirmed in replicon-containing Huh-7 cells by its ability to block NS3-mediated polyprotein processing (Fig. 2). Treatment with 0.0002–3.6 μ M BILN 2061 showed a dose–response inhibition of *cis*-cleavage occurring at the NS3–NS4A junction associated with a reduction of the mature NS3 protein and an increase in the NS3–NS5B precursor. The accumulation of NS3–NS5B precursor in the presence of BILN 2061 suggested that polyprotein processing is inhibited at all NS3-dependent cleavage sites.

BILN 2061 was evaluated in a broad spectrum of pharmacological tests in order to assess its overall pre-clinical safety profile (Supplementary Information 2). These results showed that BILN 2061 is well tolerated in a broad safety pharmacology screen composed of various *in vitro* and *in vivo* assays. The pharmacokinetic and *in vitro* metabolic properties of BILN 2061 were determined in various animal species (Supplementary Information 3). BILN 2061 showed a low to moderate oral biological availability after single-dose pharmacokinetic studies, and was slowly metabolized by hepatic microsomes in all species tested.

In a randomized, double-blind, single-dose escalation study with placebo controls BILN 2061 was administered to healthy volunteers over a range of 5 to 2,400 mg (13 dose levels or placebo) as a drinking solution in polyethylene glycol 400 (PEG 400) and ethanol (20% w/w). BILN 2061 was well tolerated up to 2,000 mg but unspecific intestinal adverse events were observed at the highest dose (2,400 mg), probably due to local gastrointestinal irritation

caused by the large drug amount. No serious clinical or laboratory findings were identified. This included the absence of any change in liver function tests at all doses. The complete safety data set of this clinical study has been reported²³ and more details will be published elsewhere (H.N., G.S. and C.-L.Y., manuscript in preparation). Pharmacokinetics of BILN 2061 in humans after oral administration demonstrated an initial rise followed by a biphasic decline in plasma concentrations (Fig. 3). The C_{max} (maximum concentration in plasma) occurred mostly within 2–4 h after administration and the mean elimination half-life was around 4 h for all dose groups. C_{max} and $AUC_{0-\infty}$ (area under the plasma concentration–time curve) appear to be dose-proportional up to 1,200 mg. From data extrapolation of the 200 mg dose group, a steady-state concentration of 42 nM could be predicted as the trough plasma concentration ($C_{min} = C_{12h}$) in a twice-daily chronic dosing regimen. The predicted C_{min} corresponds to 14-fold the cellular efficacy of BILN 2061 (EC_{50} of 3 nM). The favourable oral pharmacokinetic profile of BILN 2061 in humans and the lack of relevant adverse events in this study further support the evaluation of BILN 2061 in HCV-infected patients.

In a randomized, double-blind, proof-of-concept study with placebo controls BILN 2061 was investigated in patients infected with HCV genotype 1 in a two-day, twice-daily treatment. The plasma HCV RNA virus load was measured up to 11 days after administration in patients treated orally with 200 mg BILN 2061 or placebo (Fig. 4). BILN 2061 was highly effective, inducing a rapid decline in virus load in all treated patients (geometric mean), and reaching in some patients (Pt1 and Pt2) undetectable levels within 24–28 h after administration. This substantial and impressive effect corresponds to a 2–3 $\log_{10}$ or greater reduction in virus load for all patients treated with BILN 2061. The virus load was undetectable in most of the patients at 48 h after initiation, although it was positive at the detection limit of 50 HCV RNA copies ml^{-1} using a qualitative transcription-mediated assay (Bayer). Finally the virus load decline was followed by a virus rebound in all patients that returned to pre-treatment levels within 6–13 days after initiation of BILN 2061 treatment. No significant reduction in virus load was observed in plasma samples of placebo-treated patients. Similar substantial virus load declines were observed in BILN 2061-treated patients that were either naive to the treatment or previously treated

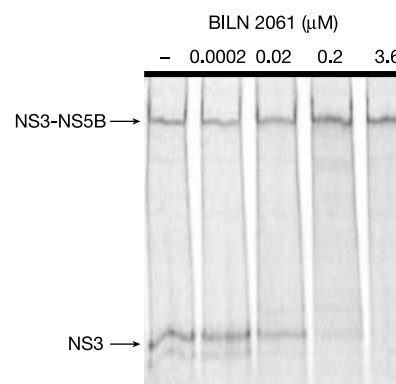


Figure 2 BILN 2061 inhibition of HCV polyprotein processing. Inhibition of NS3-protease-mediated polyprotein processing in cells containing an HCV 1b subgenomic NS2–NS5B replicon. To detect HCV non-structural protein precursors and NS3 mature protein, replicon-containing cells treated with increasing amounts of BILN 2061 were pulse-labelled for 2 h with ^{35}S -labelled methionine/cysteine. After incubation, cell extracts were immunoprecipitated with a specific anti-NS3 protein antibody and products were analysed by SDS–PAGE followed by phosphor imaging as previously described¹⁸. The position of the NS3–NS5B precursor and the NS3 mature protein are indicated.

Table 1 Biological profile of BILN 2061

Assay type/secondary activity	Inhibition values
NS3–NS4A protease assays	
HCV 1b enzyme	$K_i = 0.66$ nM
HCV 1a enzyme	$K_i = 0.30$ nM
Mechanism	Competitive and reversible
Surrogate cell-based assays	
HCV replicon 1b	$EC_{50} = 3$ nM
HCV replicon 1a	$EC_{50} = 4$ nM
Serum shift assay (50% human serum)	<Tenfold increase in EC_{50}
Cytotoxicity MTT assay	$CC_{50} = 33$ μ M
Secondary activity	
Human leukocyte elastase	$IC_{50} > 30$ μ M
Cathepsin B	$IC_{50} > 30$ μ M

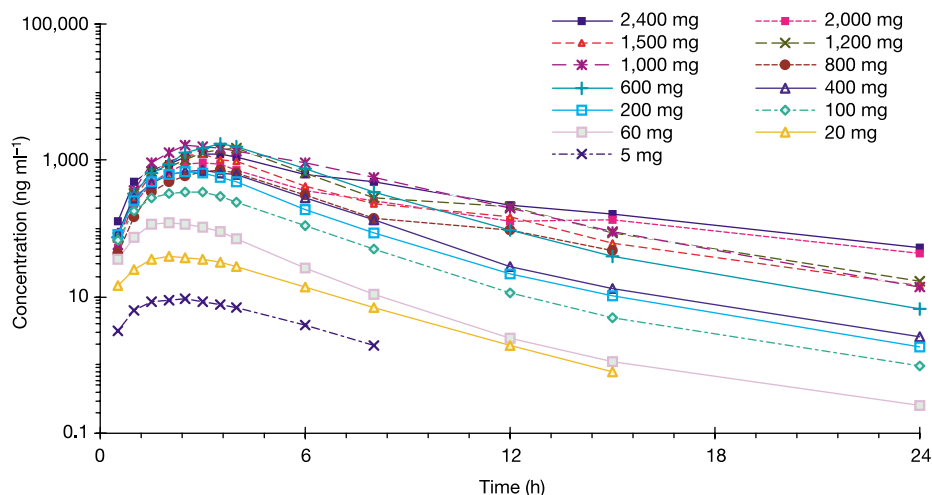


Figure 3 BILN 2061 concentration in human plasma after single-dose oral administration. BILN 2061 was exposed to healthy volunteers in a randomized, double-blind, escalating single-dose study with placebo controls. The pharmacokinetic profiles of the compound in

six active subjects were determined at the various dose levels tested. Geometric mean plasma BILN 2061 concentration versus time profiles are depicted for the various dose levels.

with interferon (IFN), as well as those with either minimal or advanced liver disease^{24,25}. A detailed description of the antiviral effect, safety and tolerability of BILN 2061 will be reported elsewhere.

The efficacy of BILN 2061 in humans establishes proof-of-concept for an NS3 protease inhibitor and a new class of a selective anti-HCV agent that was specifically designed to inhibit an essential viral enzyme. The purpose of this short trial in humans was to assess the *in vivo* antiviral efficacy of BILN 2061, which was selected only on the basis of its *in vitro* potency in surrogate enzymatic and cell culture assays and its oral pharmacokinetic profile in animals. The pharmacodynamic effect of BILN 2061 was assessed with an experimental treatment of HCV-infected humans following a trial design in which a significant virus load reduction was expected

based on current IFN-based therapies. In patients treated with BILN 2061 the extent of viral decline is significantly greater than that observed for IFN-treated patients and can be explained by a greater effectiveness of BILN 2061 in specifically blocking the production of HCV virions. Hence, the exceptional efficacy of BILN 2061 (200 mg dose) was evident by the suppression of virus load below the detection level ($1,500 \text{ RNA copies ml}^{-1}$) after only one day of treatment. In IFN-treated patients, the kinetics in patients that responded to the treatment is at best characterized by an initial virus load decline of $0.5\text{--}2 \log_{10}$ within 48 h, and required a 2–4-week treatment with either standard or pegylated IFN to attain a $3 \log_{10}$ decline in virus load^{26–28}. It has been reported recently that HCV protease inhibition may lead to a restoration of the cellular antiviral response mediated by IFN regulatory factor 3 (ref. 29) in addition to

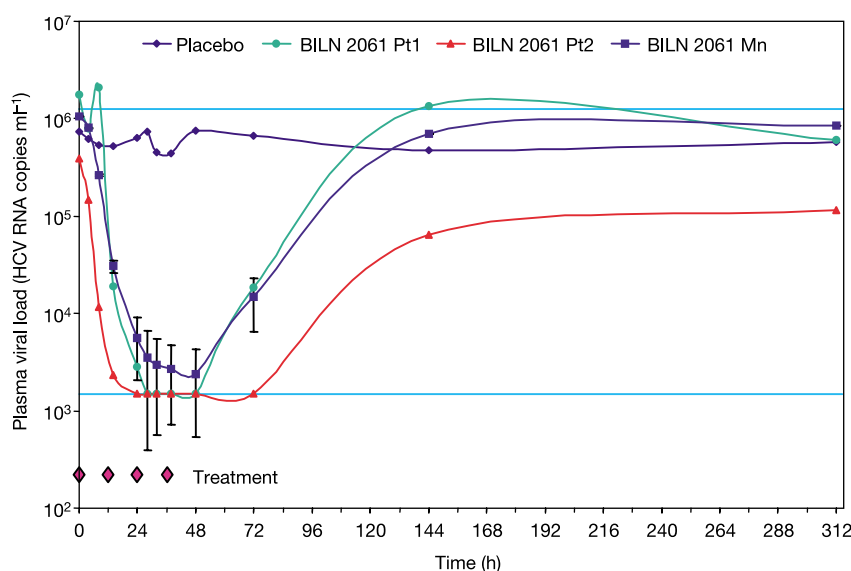


Figure 4 Antiviral efficacy of BILN 2061 in HCV-infected patients. Plasma virus load of individual patients treated with BILN 2061 (BILN 2061 Pt1, BILN 2061 Pt2), placebo (placebo) and geometric mean (BILN 2061 Mn) are shown, with standard deviation of eight patients treated with 200 mg of BILN 2061 twice daily for 2 days as an oral solution

in a PEG 400:ethanol mixture. Diamonds represent time of administration. The linear quantitative range of $1,500\text{--}1,250,000 \text{ HCV RNA copies ml}^{-1}$ of the Cobas Amplicor HCV Monitor V2.0 is indicated with horizontal blue lines.

its effects on viral replication. We cannot exclude that this may be a contributing factor in the rapid and marked virus load decline observed in patients treated with BILN 2061.

The antiviral results of protease inhibitor BILN 2061 in a proof-of-concept human trial clearly demonstrate the great potential of selective and potent anti-HCV agents. BILN 2061 will require longer trials to assess sustained antiviral activity and holds great promise to markedly improve treatments of chronic HCV infection. □

Methods

Clinical trials, ethical conduct and consent

Clinical trials were initiated after the protocol, informed consent and subject information form had been reviewed and received approval from the local Institutional Review Board (IRB) or an Independent Ethics Committee (IEC). The IRB or IEC have performed all duties outlined by the requirements of the participating countries. The trial was carried out in accordance with the principles stated in the Declaration of Helsinki and its amendments (revised version from 1996) and in accordance with Good Clinical Practice and local laws. Data generated as a result of this trial are available on request from the participating physicians, the sponsor's monitors, the quality assurance auditors, by the IRB or IEC, and the regulatory health authorities. Before subject participation in the trial, written informed consent was obtained from each subject according to the regulatory and legal requirements of the participating country.

In vitro inhibitory potency of BILN 2061

Inhibition studies were performed as previously described¹⁸ except for the use of the fluorogenic substrate anthranilyl-Asp-(D)Glu-Ile-Val-Pro-NVal[C(O)-O]-Ala-Met-Tyr(3-NO₂)-Thr-Trp-OH. *In vitro* specificity assays were performed as previously described²² using the representative serine (human leukocyte elastase) and cysteine (human liver cathepsin B) protease. For the dose-dependent inhibition of subgenomic HCV RNA levels, HCV-specific RNA copy number was quantified as previously described¹⁸ by quantitative real-time polymerase chain reaction with reverse transcription with the ABI PRISM 7700 sequence detection system, and normalized to the total cellular RNA recovered as quantified with RiboGreen (Molecular Probes) using a HCV bicistronic NS2-NS5B subgenomic replicon 1a (G.K. *et al.*, unpublished observation) and 1b corresponding to the described clone 1377/NS2-3' wt⁹. In the serum shift assay, inhibitory activity of BILN 2061 was determined using replicon 1b in the presence of 50% extracellular human serum. In the presence of BILN 2061 the percentage of inhibition was determined by reduction in HCV RNA levels, which is expressed as genome equivalents per microgram of total cellular RNA recovered relative to a DMSO control without compound. The percentage of inhibition is then plotted against the compound concentration and a nonlinear curve was fitted (Hill model) to the percentage inhibition-concentration data. The calculated percentage inhibition values were then used to determine EC₅₀ using a nonlinear regression routine procedure of SAS and the following equation: inhibition (%) = $I_{\max}[\text{inhibitor}]^n / ([\text{inhibitor}]^n + \text{EC}_{50}^n)$.

The cytotoxicity of BILN 2061 (CC₅₀) was determined in Huh-7 cells using a MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide] metabolic assay as previously described¹⁸.

Phase I and proof-of-concept clinical studies

In a randomized, double-blind, escalating single-dose study with placebo controls BILN 2061 was administered to eight male subjects (six active and two placebo) in a 10 ml PEG 400:ethanol (80%:20%) drinking solution per dose level taken on an empty stomach. Administrations were given orally as a single morning dose. The pharmacokinetic profiles of the compound in six active subjects were determined at the various dose levels tested: 5, 20, 60, 100, 200, 400, 600, 800, 1,000, 1,200, 1,500, 2,000 and 2,400 mg. Plasma samples obtained at various times were analysed by a liquid chromatography and mass spectrometry method. The specific procedure and specification of the analytical methods will be reported elsewhere (C.-L.Y. *et al.*, unpublished observation). Pharmacokinetic parameters (Supplementary Information 3) were calculated using a non-compartmental method. Dose proportionality of plasma BILN 2061 concentrations in terms of A_{0-∞} and C_{max} was assessed by regression analysis.

In a randomized, double-blind, proof-of-concept study with placebo controls the antiviral activity of BILN 2061 was investigated in ten patients infected with HCV genotype 1 (INNO-LiPA HCV II, Innogenetics) and with minimal liver fibrosis as determined histologically. Patients were treated with 200 mg BILN 2061 (*n* = 8) or placebo twice daily for 2 days as an oral solution in a PEG 400:ethanol (80:20 w/w) mixture. Plasma samples were drawn at various times and plasma HCV RNA levels were determined using the Cobas Amplicor HCV Monitor V2.0 (Roche Diagnostics), which has a limit of detection of 1,500 HCV RNA copies ml⁻¹ and a linear quantitative range from 1,500 to 1,250,000 HCV RNA copies ml⁻¹. All samples were analysed by a central laboratory. Plasma samples were also subjected to the branched DNA assay (Bayer), resulting in a linear quantitative measurement from 1,500 to 40,000,000 HCV RNA copies ml⁻¹, and to the transcription-mediated amplification assay (Bayer), resulting in a lower range assay sensitivity of 50 HCV RNA copies ml⁻¹.

Received 25 August; accepted 2 October 2003; doi:10.1038/nature02099.

Published online 26 October 2003.

1. Choo, Q. L. *et al.* Isolation of a cDNA clone derived from a blood-borne non-A non-B viral hepatitis genome. *Science* **244**, 359–362 (1989).
2. Kuo, G. *et al.* An assay for circulating antibodies to a major etiologic virus of human non-A, non-B

hepatitis. *Science* **244**, 362–364 (1989).

3. Di Bisceglie, A. M. Hepatitis C. *Lancet* **351**, 351–355 (1998).
4. Chander, G. *et al.* Treatment of chronic hepatitis C: a systematic review. *Hepatology* **36**, S135 (2002).
5. Tan, S.-L., Pause, A., Shi, Y. & Sonenberg, N. Hepatitis C therapeutics: current status and emerging strategies. *Nature Rev. Drug Discov.* **1**, 867–881 (2002).
6. Di Bisceglie, A. M., McHutchison, J. & Rice, C. M. New therapeutic strategies for hepatitis C. *Hepatology* **35**, 224–231 (2002).
7. Reed, K. E. & Rice, C. M. Overview of hepatitis C virus genome structure, polyprotein processing, and protein properties. *Curr. Top. Microbiol. Immunol.* **242**, 55–84 (2000).
8. Kolykhalov, A. A., Mihalik, K., Feinstone, S. M. & Rice, C. M. Hepatitis C virus-encoded enzymatic activities and conserved RNA elements in the 3' nontranslated region are essential for virus replication *in vivo*. *J. Virol.* **74**, 2046–2051 (2000).
9. Lohmann, V. *et al.* Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. *Science* **285**, 103–107 (1999).
10. Mercer, D. F. *et al.* Hepatitis C virus replication in mice with chimeric human livers. *Nature Med.* **7**, 927–933 (2001).
11. Schechter, I. & Berger, A. Protease substrate nomenclature. *Biochem. Biophys. Res. Commun.* **27**, 157–162 (1967).
12. Llinàs-Brunet, M. *et al.* Peptide-based inhibitors of the hepatitis C virus serine protease. *Bioorg. Med. Chem. Lett.* **8**, 1713–1718 (1998).
13. Steinkühler, C. *et al.* Product inhibition of the hepatitis C virus NS3 protease. *Biochemistry* **37**, 8899–8905 (1998).
14. Llinàs-Brunet, M. *et al.* Studies on the C-terminal of hexapeptide inhibitors of the hepatitis C virus serine protease. *Bioorg. Med. Chem. Lett.* **8**, 2719–2724 (1998).
15. Llinàs-Brunet, M. *et al.* Highly potent and selective peptide-based inhibitors of the hepatitis C virus serine protease: towards smaller inhibitors. *Bioorg. Med. Chem. Lett.* **10**, 2267–2270 (2000).
16. Poupard, M.-A. *et al.* Solid-phase synthesis of peptidomimetic inhibitors for the hepatitis C virus NS3 protease. *J. Org. Chem.* **66**, 4743–4751 (2001).
17. Goudreau, N. *et al.* NMR structural characterization of peptide inhibitors bound to the HCV NS3 protease: design of a new P2 substituent. *J. Med. Chem.* (submitted).
18. Pause, A. *et al.* An NS3 serine protease inhibitor abrogates replication of subgenomic hepatitis C virus RNA. *J. Biol. Chem.* **278**, 20374–20380 (2003).
19. LaPlante, S. R. *et al.* Solution structure of substrate-based ligands when bound to hepatitis C virus NS3 protease domain. *J. Biol. Chem.* **274**, 18618–18624 (1999).
20. LaPlante, S. R. *et al.* NMR line-broadening and transferred NOESY as a medicinal chemistry tool for studying inhibitors of the hepatitis C virus NS3 protease domain. *Bioorg. Med. Chem. Lett.* **10**, 2271–2274 (2000).
21. Boehringer Ingelheim (Canada) Ltd. Hepatitis C inhibitor tripeptides. US patent 6,323,180 B1 (2001).
22. Tsantirizos, Y. *et al.* Macrocyclic inhibitors of the NS3 protease as potential therapeutic agents of hepatitis C virus infection. *Angew. Chem. Int. Edn Engl.* **42**, 1356–1360 (2003).
23. Narjes, H., Yang, C. L., Stahle, H. & Steinmann, G. Tolerability and pharmacokinetics of BILN 2061: a novel serine protease HCV inhibitor after oral single doses of 5 mg to 2,400 mg in healthy male subjects. *Hepatology* **36**(4), Abstr. 800 (2002).
24. Hinrichsen, H. *et al.* First report on the antiviral efficacy of BILN 2061, a novel oral serine protease inhibitor, in patients with chronic hepatitis C genotype 1. *Hepatology* **36**, 297A, Abstr. 866 (2002).
25. Benhamou, Y. *et al.* Safety, tolerability and antiviral effect of BILN 2061, a novel HCV serine protease inhibitor, after oral treatment over 2 days in patients with chronic hepatitis C, genotype 1, with advanced liver fibrosis. *Hepatology* **36**, 304A, Abstr. 563 (2002).
26. Neumann, A. U. *et al.* Hepatitis C viral dynamics *in vivo* and the antiviral efficacy of interferon-α therapy. *Science* **282**, 103–107 (1998).
27. Neumann, A. U. *et al.* Differences in viral dynamics between genotypes 1 and 2 of hepatitis C virus. *J. Infect. Dis.* **182**, 28–35 (2000).
28. Zeuzem, S. *et al.* Viral kinetics in patients with chronic hepatitis C treated with standard or peginterferon α 2a. *Gastroenterology* **120**, 1438–1447 (2001).
29. Foy, E. *et al.* Regulation of interferon regulatory factor-3 by the hepatitis C virus serine protease. *Science* **300**, 1145–1148 (2003).
30. Boehringer Ingelheim (Canada) Ltd. Macrocyclic peptides active against the Hepatitis C virus. US patent 6,608,027 B1 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We wish to thank F. Dô, J. Duan, M. Garneau, N. Lapeyre-Paquette, S. Lefebvre, F. Liard, M. Marquis, R. Maurice, G. McKercher, A. Pause, C. Pellerin, L. Pilote, C. Plouffe and M. Rhéaume from the Biological Sciences Department, and N. Aubry, J. Bordeleau, C. Boucher, Y. Bousquet, C. Brochu, C. Chabot, R. Déziel, J.-S. Duceppe, G. Fazal, J.-M. Ferland, E. Ghirio, J. Gillard, V. Gorys, S. Goulet, C. Grand-Maitre, B. Haché, T. Halmos, P. Lavallée, M. Little, E. Malenfant, R. Plante, M. Poirier, S. Valois and D. Wernic from the Chemistry Department. We also thank B. Willems and the Centre de Recherche Clinique André Viallet, Hôpital St-Luc, CHUM. We acknowledge the BILN 2061 development team and members of the departments of Drug Metabolism and Pharmacokinetics, Pharmacology, Pharmaceuticals, Analytical Sciences, Process Chemistry, Toxicology, Drug Regulatory Affairs and Medical at Boehringer Ingelheim Pharmaceuticals, Inc., Connecticut, USA; J. Croenlein, G. Nehmiz and the clinical research teams at Boehringer Ingelheim Pharma KG, Biberach, Germany; and Academisch Medisch Centrum, The Netherlands. We thank the volunteers and patients who participated in the clinical trials of BILN 2061. R.E.S. was clinical investigator and received honorarium from Boehringer Ingelheim. All other authors are or were employees of Boehringer Ingelheim.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.L. (dlamarre@lav.boehringer-ingelheim.com).

CREB controls hepatic lipid metabolism through nuclear hormone receptor PPAR- γ

Stephan Herzig¹, Susan Hedrick¹, Ianessa Morantte¹, Seung-Hoi Koo¹, Francesco Galimi² & Marc Montminy¹

¹Peptide Biology Laboratories and ²Laboratory of Genetics, Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037-1002, USA

Fasting triggers a series of hormonal cues that promote energy balance by inducing glucose output and lipid breakdown in the liver¹. In response to pancreatic glucagon and adrenal cortisol, the cAMP-responsive transcription factor CREB activates gluconeogenic and fatty acid oxidation programmes by stimulating expression of the nuclear hormone receptor coactivator PGC-1 (refs 2–5). In parallel, fasting also suppresses lipid storage and synthesis (lipogenic) pathways¹, but the underlying mechanism is unknown. Here we show that mice deficient in CREB activity have a fatty liver phenotype and display elevated expression of the nuclear hormone receptor PPAR- γ , a key regulator of lipogenic genes^{6,7}. CREB inhibits hepatic PPAR- γ expression in the fasted state by stimulating the expression of the Hairy Enhancer of Split (*HES-1*) gene, a transcriptional repressor that is shown here to be a mediator of fasting lipid metabolism *in vivo*. The coordinate induction of PGC-1 and repression of PPAR- γ by CREB during fasting provides a molecular rationale for the antagonism between insulin and counter-regulatory hormones, and indicates a potential role for CREB antagonists as therapeutic agents in enhancing insulin sensitivity in the liver.

The ability of CREB (CRE-binding protein) to act as a prominent checkpoint in hepatic gluconeogenesis prompted us to examine the importance of this factor for hepatic lipid metabolism during fasting. When compared with control littermates, mice infected with a dominant-negative CREB-expressing adenovirus (ACREB) showed a fatty liver phenotype and a pronounced increase in hepatic triglyceride content (49 ± 2.1 versus 18 ± 1.4 mg per g tissue) and in plasma triglyceride levels (58 ± 2 versus 42 ± 5 mg dl⁻¹) on a high-fat diet (Fig. 1a; Supplementary Fig. S1). CREB^{+/-} heterozygotes⁸ also displayed higher liver triglyceride contents than wild-type littermates (12 ± 0.6 versus 9 ± 0.3 mg per g tissue; Supplementary Fig. S2). Contrary to its inhibitory effects on gluconeogenic and fatty acid oxidation genes during fasting (Supplementary Fig. S3), ACREB stimulated the hepatic expression of lipogenic genes such as fatty acid transporter CD36, fatty acid synthase (FAS), ATP citrate lyase (ACL) and caveolin (CAV) twofold to threefold over control littermates (Fig. 1b). Arguing against nonspecific effects of ACREB on this programme, knock down of CREB expression using CREB RNA interference (RNAi) adenovirus also induced lipogenic gene expression in primary rat hepatocytes (Supplementary Fig. S4). Neither ACREB nor CREB RNAi altered hepatic levels of the sterol-regulatory-element-binding protein SREBP-1c, which is a master regulator of lipogenesis⁹; this argues against SREBP as a target for CREB activity in this tissue (Fig. 1b; Supplementary Fig. S4).

The elevated expression of the lipogenic programme in the livers of CREB-deficient mice suggested that the cAMP pathway normally suppresses these genes under fasting conditions. However, the absence of identifiable cAMP-responsive elements (CREs) in lipogenic genes¹⁰, favoured an indirect mechanism for CREB inhibition. While performing gene-profiling studies, we noticed that messenger

RNA and protein levels of the nuclear hormone receptor PPAR- γ (peroxisome-proliferator-activated receptor- γ), a key regulator of lipogenic genes, were induced fourfold to fivefold over control littermates in the livers of ACREB-infected and in CREB^{+/-} mice, as well as in isolated primary hepatocytes infected with CREB RNAi adenovirus (Fig. 1c; Supplementary Figs S4 and S5; data not shown).

To determine whether PPAR- γ contributes to the fat phenotype in CREB-deficient mice, we used a PPAR- γ RNAi adenovirus. Consistent with its ability to lower PPAR- γ protein levels in liver (Fig. 2a), PPAR- γ RNAi adenovirus reduced triglyceride levels (Fig. 2b) and lipogenic gene expression (Fig. 2c) in mice co-infected with ACREB adenovirus. Levels of the ACREB inhibitor protein were identical in mice injected with ACREB plus PPAR- γ RNAi viruses compared to mice injected with ACREB adenovirus alone, arguing against a nonspecific effect of PPAR- γ RNAi on ACREB expression (data not shown). These results indicate that CREB inhibits hepatic triglyceride synthesis and storage during fasting through the repression of PPAR- γ .

Similar to other lipogenic genes, the PPAR- γ promoter lacks a consensus CRE¹⁰, suggesting that CREB regulates this gene through an indirect mechanism. The importance of E-box motifs for activation of the PPAR- γ promoter by SREBP-1/ADD1 (refs 11, 12) during feeding prompted us to investigate the importance of

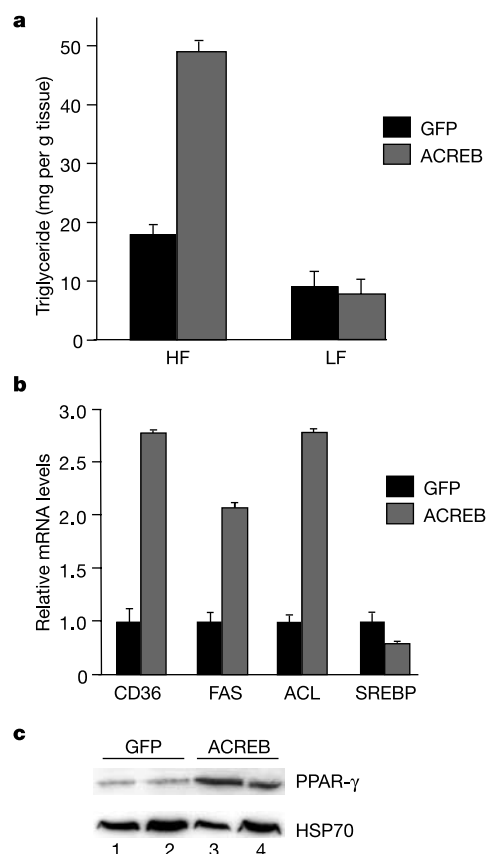


Figure 1 Mice deficient in hepatic CREB activity display a fatty liver phenotype. **a**, Fasting liver triglyceride levels in mice (mean $\pm$ s.e.m., $n = 5$ per group) injected with either control (GFP) or ACREB adenovirus. Animals were maintained on regular low-fat (LF) or high-fat (2% cholesterol; HF) diet. **b**, Quantitative PCR analysis of lipogenic (CD36, FAS, ACL and SREBP) mRNAs in fasted livers of ACREB- or GFP-expressing mice. **c**, Western blot of liver extracts from two mice injected with GFP (lanes 1 and 2) or ACREB (lanes 3 and 4) detected using PPAR- γ or HSP70 antibodies.

these regulatory elements for repression of the PPAR- γ promoter by the cAMP-CREB pathway. In transient assays of HepG2 hepatoma cells, protein kinase A (PKA) potentially inhibited PPAR- γ 1 promoter activity through sequences from -1,500 to -160 that contain six E-box elements; deletion of these sequences abolished the PKA effect completely (Fig. 2d).

The bHLH (basic helix-loop-helix) family of E-box binding proteins contains a group of transcriptional repressors that includes Hairy Enhancer of Split (HES-1), DEC1 and CLOCK¹³⁻¹⁵. In gene-profiling assays, HES-1 expression was inhibited threefold to four-

fold in livers of ACREB-infected mice relative to control mice, but DEC1 and CLOCK mRNA levels were comparable in both groups (Fig. 3a). Consistent with the notion that CREB specifically regulates the expression of HES-1, protein levels for this repressor were increased fourfold in cultured hepatocytes in response to forskolin treatment (Fig. 3a, inset). Inspection of the human *HES-1* promoter revealed a consensus CRE half-site (TGACG) at -215 that is conserved in the murine *HES-1* gene. In transient assays of HepG2 hepatoma cells, PKA induced *HES-1* promoter activity eightfold, and this effect was blocked by co-expression of ACREB inhibitor (Fig. 3b). Indeed, in chromatin immunoprecipitation (ChIP) studies, the *HES-1* promoter was efficiently recovered from immunoprecipitates using CREB antibodies but not of control immunoglobulin- γ (IgG; Fig. 3c), showing that HES-1 is a direct target for CREB action in hepatocytes.

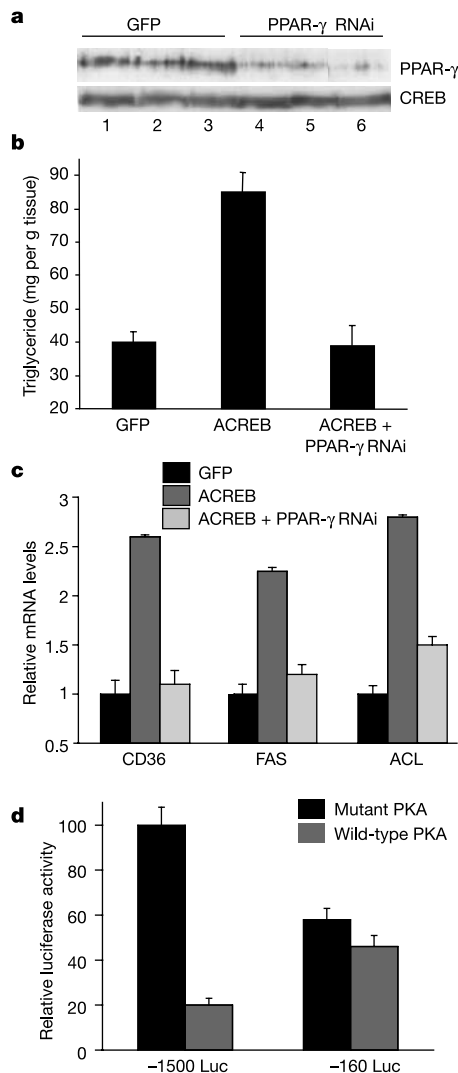


Figure 2 PPAR- γ mediates effects of CREB on triglyceride metabolism in liver. **a**, Western blot assay of liver extracts showing PPAR- γ levels in mice injected with GFP (lanes 1–3) or PPAR- γ RNAi (lanes 4–6) adenovirus. Levels of CREB protein shown as control for nonspecific effects of viral infection. Western blot of liver extracts from mice injected with GFP (lanes 1–3) or PPAR- γ RNAi (lanes 4–6) adenovirus detected using PPAR- γ - or CREB-specific antibodies. Mice were maintained on a high-fat diet. **b**, Knock down of PPAR- γ expression corrects lipid metabolism in ACREB mice. Left, triglyceride levels in liver extracts of fasted mice (mean $\pm$ s.e.m., $n = 6$ per group) injected with either GFP, ACREB or ACREB plus PPAR- γ RNAi adenovirus. **c**, Effect of PPAR- γ RNAi adenovirus on expression of lipogenic genes (CD36, FAS and ACL) in mice co-injected with ACREB adenovirus (ACREB plus PPAR- γ) versus GFP control or ACREB adenovirus alone. **d**, Transient assay of HepG2 hepatocytes co-transfected with PPAR- γ 1 reporters (-1,500 Luc or -160 Luc) and mutant or wild-type PKA expression vector.

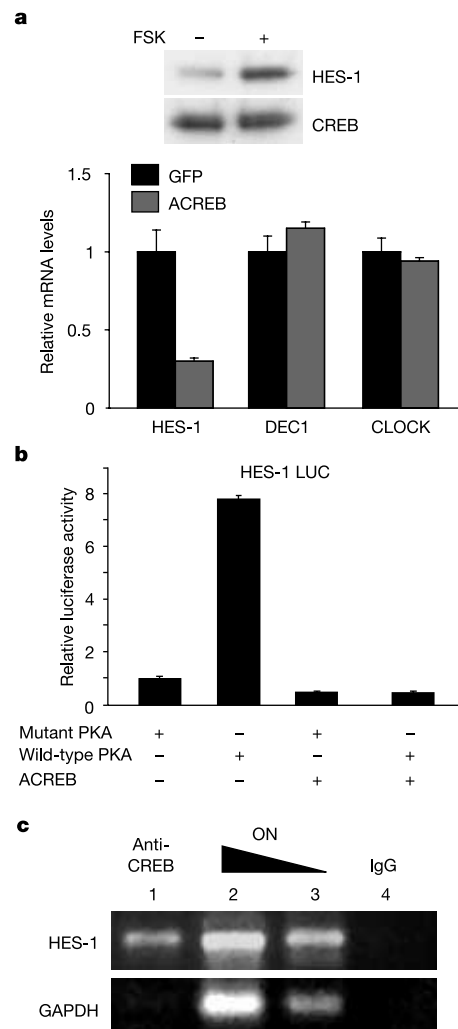


Figure 3 HES-1 is a direct CREB target *in vivo*. **a**, Effect of ACREB on fasting levels of HES-1, DEC1 and CLOCK mRNAs in livers of mice injected with control GFP or ACREB adenovirus (mean $\pm$ s.e.m., $n = 3$ per group). Inset, western blot of HES-1 and CREB levels in extracts from H4IIE cells treated for 1 h with forskolin (FSK) as indicated. **b**, Transient assay of HepG2 hepatocytes transfected with HES-1 reporter (HES-1 Luc). Co-transfection with mutant or wild-type PKA and ACREB constructs as indicated. **c**, ChIP assay of *HES-1* promoter or *GAPDH* gene in HepG2 cells using CREB or control IgG antisera. Lanes 2 and 3, 1% and 0.1% DNA inputs, respectively.

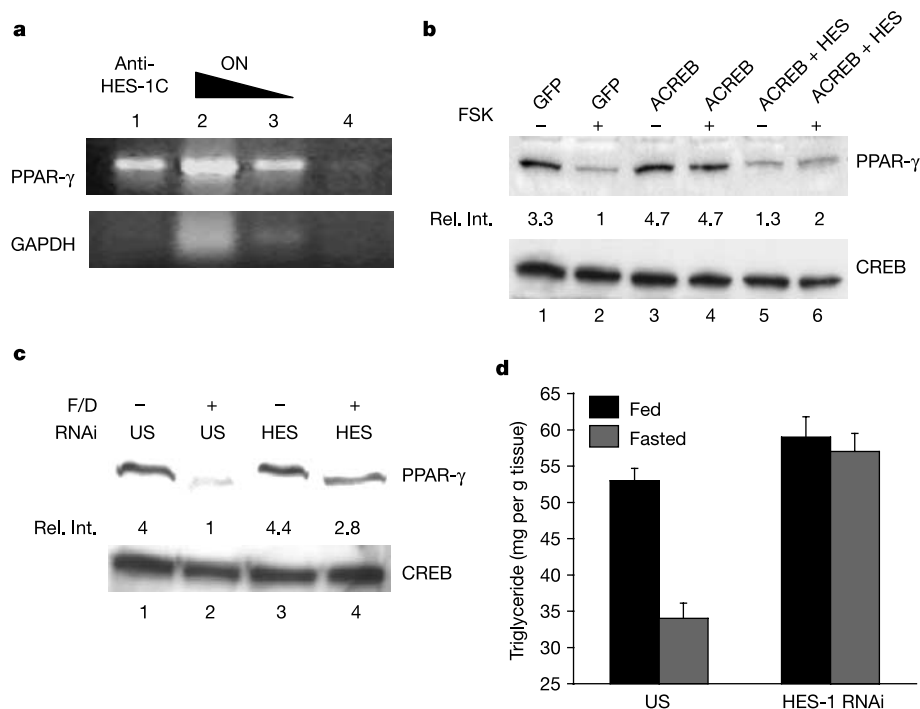


Figure 4 HES-1 mediates CREB-dependent inhibition of the lipogenic pathway during fasting. **a**, ChIP assay of *PPARG1* promoter and control *GAPDH* gene in HepG2 cells transfected with a Flag-tagged HES-1 (lane 1) or empty Flag (C; lane 4) construct using anti-Flag antibody. Lanes 2 and 3, 1% and 0.1% inputs. **b**, Western blot showing effect of forskolin (FSK) on PPAR-γ levels in cells infected with either GFP, ACREB or ACREB plus HES-1 adenoviruses. Relative amounts of PPAR-γ (Rel. Int.) normalized to CREB protein

levels. **c**, Western blot of PPAR-γ and CREB levels in H4IIE cells infected with HES-1 or control (US) RNAi adenoviruses. Overnight treatment with forskolin (10 μM) and dexamethasone (1 μM) (F/D) indicated. **d**, Hepatic triglyceride levels in mice injected with either control (US) or HES-1 RNAi adenovirus under refed or fasted conditions (mean ± s.e.m., *n* = 5 per group). Mice were maintained on a high-fat diet.

HES-1 preferentially binds to two subclasses of E-box motifs, referred to as the B (CANGTG) and N (CACNAG) classes¹⁶. In affinity selection assays using E boxes within the PPAR-γ1 promoter, HES-1 was found to bind with highest affinity to a B class E box at -790, with weaker binding at two N boxes (-250 and -700) (Supplementary Fig. S6). In ChIP assays, we detected the endogenous PPAR-γ1 promoter in HES-1 immunoprecipitates prepared from HepG2 hepatocytes transiently expressing a Flag-tagged HES-1 transgene at comparable levels to endogenous HES-1 (Fig. 4a). These results suggest that the *PPARG1* gene is indeed a target for HES-1 in the liver.

To test the role of HES-1 in regulating PPAR-γ expression, we used a HES-1-expressing adenovirus. Treatment with a cAMP agonist reduced PPAR-γ protein levels 3.3-fold in hepatoma cells, and this effect was reversed by expression of ACREB inhibitor (Fig. 4b, lanes 2 and 4). Overexpression of HES-1 fully repressed PPAR-γ even in the presence of the ACREB inhibitor, showing that HES-1 acts downstream of CREB (Fig. 4b, lanes 4 and 6).

To test further the importance of HES-1 as a PPAR-γ repressor, we developed a HES-1 RNAi adenovirus. Consistent with its ability to knock down expression of endogenous HES-1 protein in hepatoma cells (Supplementary Fig. S7), HES-1 RNAi adenovirus blocked the inhibitory effects of cAMP and dexamethasone treatment on PPAR-γ1 promoter activity (Supplementary Fig. S8), as well as on PPAR-γ protein (Fig. 4c, lanes 2 and 4). These results demonstrate that endogenous HES-1 mediates the inhibitory effects of the cAMP-CREB pathway on PPAR-γ expression in hepatocytes.

To determine the importance of HES-1 for fasting lipid metabolism *in vivo*, we used HES-1 RNAi adenovirus to knock down hepatic HES-1 protein expression in mice. Remarkably, fasting

hepatic triglyceride levels were abnormally elevated (HES-1 RNAi, 57 ± 5 mg per g tissue; control, 34 ± 6 mg per g tissue) in HES-1-deficient animals, recapitulating observations with ACREB-infected and CREB^{+/-} mice (Fig. 4d). Consistent with these changes, amounts of PPAR-γ, CAV and ACL mRNAs in liver were also higher during fasting in HES-1-deficient mice compared with control littermates (Supplementary Fig. S9). These results support the notion that HES-1 mediates the inhibitory effects of CREB on the lipogenic pathway under fasting conditions.

Our studies provide a novel example of crosstalk between cell-surface and nuclear hormone receptor pathways. The ability of CREB to coordinate hepatic lipid and glucose metabolism during fasting through inhibition of PPAR-γ and induction of PGC-1 (PPAR-γ coactivator 1), respectively, indicates that CREB may contribute to insulin resistance in susceptible individuals. In this regard, the development of CREB antagonists that block recruitment of the CREB co-activator CBP (CREB-binding protein) in liver may provide a useful therapy for diabetic patients by enhancing insulin sensitivity and glycaemic control. □

Methods

Recombinant adenoviruses

Adenoviruses expressing ACREB or green fluorescent protein (GFP) have been described previously². RNAi adenoviruses contained CREB-, PPAR-γ- or HES-1-specific oligonucleotides under control of the U6 promoter¹⁷. Viruses were purified by the caesium chloride method and dialysed against PBS buffer containing 10% glycerol before animal injections as described².

Animal experiments

Male 5-week-old C57Bl6 mice were obtained from The Jackson Laboratory and maintained on a 12-h light-dark cycle with regular unrestricted diet. For virus injections, animals were anaesthetized with Iso-Flurane, and a total of 1 × 10⁹ plaque-forming units per recombinant virus was administered by systemic tail vein injection. In each

experiment, at least five animals received identical treatments. Effects of ACREB and control GFP adenoviruses on metabolic parameters and liver gene expression were measured 5 days after injection. Animals were fasted for 18 h overnight with free access to water, or fasted for 18 h and refed for the following 24 h. Additional liver samples were fixed in 10% formaldehyde, sectioned with a cryomicrotome, and investigated for viral infection efficiency by fluorescence microscopy. Studies with CREB knockout mice carrying a targeted disruption of the CREB DNA-binding domain and leucine zipper region⁸ were performed as described above. For fasting–refeeding experiments, male 5-week-old C57Bl6 wild-type mice were left with free access to food and water, fasted for 24 h overnight with free access to water, or fasted for 24 h and refed for another 24 h.

Blood metabolites

Blood glucose values were determined from whole blood using an automatic glucose monitor (One Touch, Lifescan). Plasma triglyceride, total cholesterol and free fatty acid levels were determined enzymatically using triglyceride GPO-Trinder, Infinity Cholesterol and NEFA C reagents, respectively (Sigma; Wako).

Tissue lipid extraction

Hepatic lipids were extracted as described previously¹⁸, and triglyceride, total cholesterol and free fatty acid contents were determined using the commercial kits described above. Values were calculated as mg per g wet tissue (triglycerides and cholesterol) or μmol per g wet tissue (free fatty acids).

Primary hepatocyte isolation

Primary hepatocytes were isolated from male Harlan Sprague–Dawley rats (180–260 g) using the collagenase perfusion method as described previously¹⁹. After a 3–6 h attachment period, cells were infected with recombinant adenovirus in serum-free medium 199 including antibiotics for 16 h. Subsequently, cells were cultured in the same medium for 32 h and collected for total RNA isolation.

Quantitative RT–PCR

Total RNA was extracted from homogenized mouse liver or from primary hepatocytes using the RNeasy kit (Qiagen). Complementary DNA (cDNA) was prepared by reverse transcription of 250 ng total RNA using Superscript II enzyme and oligo dT primer (GIBCO BRL). cDNAs were amplified using the SYBR green PCR kit and an ABI PRISM 7700 Sequence detector (Perkin Elmer). RNA expression data were normalized to levels of ribosomal 36B4 RNA.

Protein analysis

Protein was extracted from frozen liver samples in SDS–urea lysis buffer, and 20 μg protein were separated on a 4–20% gradient SDS–polyacrylamide gel (Invitrogen) and blotted onto nitrocellulose membrane. Western blot assays were performed as described²⁰. Owing to rapid turnover of HES-1 protein²¹, a proteasome inhibitor (50 μM MG132) was included in all samples (plus or minus forskolin).

Chromatin immunoprecipitation assay

Human hepatoma HepG2 cells were grown to 90% confluence, and ChIP assays were performed as described elsewhere²². Precipitated DNA fragments were analysed by polymerase chain reaction (PCR) amplification using primers directed against the human *HES-1* promoter, the human *PPARG1* promoter (5′-CAGTGGTCAGTCCCTGTAAGA TCC; 3′-GAGCATACTGTTCTGGGCCCTGATCCT), or the *GAPDH* (glyceraldehyde-3-phosphate dehydrogenase) coding region as negative control.

Avidin–biotin conjugate DNA precipitation assay

HepG2 cells were grown to 90% confluence, and avidin–biotin conjugate DNA precipitation (ABCD) assays were performed as described elsewhere²³. Wild-type and mutated N class E box of the *HES-1* promoter were used as positive and negative controls, respectively²⁴. Bound *HES-1* protein was detected by western blot using *HES-1*-specific antiserum.

RNA interference

Oligonucleotides corresponding to human *HES-1* (5′-GGCGCAGATGACGGCTGC GCA-3′) were annealed, cloned into pBS/U6 (ref. 17), and transfected into HepG2 cells (200 ng plasmid per well). The effect of RNAi on PKA/ACREB-dependent *PPARG1* promoter activity was measured after 48 h. Unspecific oligonucleotides (5′-GGCATTA CAGTATCGATCAGA-3′) were used as a control in all experiments.

Plasmids

pZEO-ACREB²⁰, mutant PKA (mut PKA), wild-type PKA (wt PKA)²⁵, *HES-1* Luc (luciferase; −467 to +46)²⁴ and pCMV2-Flag–*HES-1* (ref. 26) expression plasmids have been described previously. Reporter plasmids for −1,500 and −160 PPAR- γ 1 were constructed using standard cloning techniques.

Cell culture and transient transfection assays

HepG2 cells were transfected using the Fugene 6 reagent (Roche Applied Sciences) according to the manufacturer's instructions (50 ng indicator plasmid per well). Where indicated, expression plasmids encoding ACREB, mutant or wild-type PKA, or rat *HES-1*

were co-transfected (50 ng plasmid per well, respectively). Co-transfections were performed with a constant amount of DNA by balancing with empty pCDNA3 vector (Invitrogen). Cell extracts were prepared 48 h after transfection, and luciferase assays were performed as described previously²⁷, normalizing to activity from co-transfected β -galactosidase expression plasmid (50 ng per well).

Received 8 September; accepted 7 October 2003; doi:10.1038/nature02110.

1. Saltiel, A. R. New perspectives into the molecular pathogenesis and treatment of type 2 diabetes. *Cell* **104**, 517–529 (2001).
2. Herzig, S. *et al.* CREB regulates hepatic gluconeogenesis through the coactivator PGC-1. *Nature* **413**, 179–183 (2001).
3. Puigserver, P. *et al.* Insulin-regulated hepatic gluconeogenesis through FOXO1–PGC-1 α interaction. *Nature* **423**, 550–555 (2003).
4. Vega, R. B., Huss, J. M. & Kelly, D. P. The coactivator PGC-1 cooperates with peroxisome proliferator-activated receptor α in transcriptional control of nuclear genes encoding mitochondrial fatty acid oxidation enzymes. *Mol. Cell. Biol.* **20**, 1868–1876 (2000).
5. Yoon, J. C. *et al.* Control of hepatic gluconeogenesis through the transcriptional coactivator PGC-1. *Nature* **413**, 131–138 (2001).
6. Yu, S. *et al.* Adipocyte-specific gene expression and adipogenic steatosis in the mouse liver due to peroxisome proliferator-activated receptor γ 1 (PPAR- γ 1) overexpression. *J. Biol. Chem.* **278**, 498–505 (2003).
7. Matsusue, K. *et al.* Liver-specific disruption of PPAR γ in leptin-deficient mice improves fatty liver but aggravates diabetic phenotypes. *J. Clin. Invest.* **111**, 737–747 (2003).
8. Rudolph, D. *et al.* Impaired fetal T cell development and perinatal lethality in mice lacking the cAMP response element binding protein. *Proc. Natl Acad. Sci. USA* **95**, 4481–4486 (1998).
9. Horton, J. D. & Shimomura, I. Sterol regulatory element-binding proteins: activators of cholesterol and fatty acid biosynthesis. *Curr. Opin. Lipidol.* **10**, 143–150 (1999).
10. Conkright, M. D. *et al.* Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness. *Mol. Cell* **11**, 1101–1108 (2003).
11. Fajas, L. *et al.* The organization, promoter analysis, and expression of the human PPAR- γ gene. *J. Biol. Chem.* **272**, 18779–18789 (1997).
12. Fajas, L. *et al.* Regulation of peroxisome proliferator-activated receptor γ expression by adipocyte differentiation and determination factor 1/sterol regulatory element binding protein 1: implications for adipocyte differentiation and metabolism. *Mol. Cell. Biol.* **19**, 5495–5503 (1999).
13. Sasaki, Y., Kageyama, R., Tagawa, Y., Shigemoto, R. & Nakanishi, S. Two mammalian helix–loop–helix factors structurally related to *Drosophila* hairy and Enhancer of split. *Genes Dev.* **6**, 2620–2634 (1992).
14. Jarriault, S. *et al.* Signalling downstream of activated mammalian Notch. *Nature* **377**, 355–358 (1995).
15. Gekakis, N. *et al.* Role of the CLOCK protein in the mammalian circadian mechanism. *Science* **280**, 1564–1569 (1998).
16. Iso, T., Kedes, L. & Hamamori, Y. HES and HERP families: multiple effectors of the Notch signaling pathway. *J. Cell. Physiol.* **194**, 237–255 (2003).
17. Sui, G. *et al.* A DNA vector-based RNAi technology to suppress gene expression in mammalian cells. *Proc. Natl Acad. Sci. USA* **99**, 5515–5520 (2002).
18. Lee, Y. *et al.* Liporegulation in diet-induced obesity. The antisteatotic role of hyperleptinemia. *J. Biol. Chem.* **276**, 5629–5635 (2001).
19. Koo, S. H. & Towle, H. C. Glucose regulation of mouse S(14) gene expression in hepatocytes. Involvement of a novel transcription factor complex. *J. Biol. Chem.* **275**, 5200–5207 (2000).
20. Michael, L. F., Asahara, H., Shulman, A. I., Kraus, W. L. & Montminy, M. The phosphorylation status of a cyclic AMP-responsive activator is modulated via a chromatin-dependent mechanism. *Mol. Cell. Biol.* **20**, 1596–1603 (2000).
21. Hirata, H. *et al.* Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop. *Science* **298**, 840–843 (2002).
22. Canetti, G. *et al.* Attenuation of a phosphorylation-dependent activator by an HDAC–PP1 complex. *Nature Struct. Biol.* **10**, 175–181 (2003).
23. Daitoku, H., Yamagata, K., Matsuzaki, H., Hatta, M. & Fukamizu, A. Regulation of PGC-1 promoter activity by protein kinase B and the forkhead transcription factor FKHR. *Diabetes* **52**, 642–649 (2003).
24. Takebayashi, K. *et al.* Structure, chromosomal locus, and promoter analysis of the gene encoding the mouse helix–loop–helix factor HES-1. Negative autoregulation through the multiple N box elements. *J. Biol. Chem.* **269**, 5150–5156 (1994).
25. Orellana, S. A. & McKnight, G. S. Mutations in the catalytic subunit of cAMP-dependent protein kinase result in unregulated biological activity. *Proc. Natl Acad. Sci. USA* **89**, 4726–4730 (1992).
26. McLaren, K. W. *et al.* The mammalian basic helix loop helix protein HES-1 binds to and modulates the transactivating function of the runt-related factor Cbfa1. *J. Biol. Chem.* **275**, 530–538 (2000).
27. Nakajima, T., Uchida, C., Anderson, S. F., Parvin, J. D. & Montminy, M. Analysis of a cAMP-responsive activator reveals a two-component mechanism for transcriptional induction via signal-dependent factors. *Genes Dev.* **11**, 738–747 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank K. Suter and L. Vera for performing injections; S. Stifani, R. Kageyama, J. Auwerx, Y. Shi and T. Sudo for providing reagents; and G. Schuetz for CREB knockout mice. We also thank R. Evans for reviewing the manuscript, and I. Verma for support. This work was supported by the NIH (M.M.), the American Diabetes Association, the Hillblom Foundation and the Deutsche Forschungsgemeinschaft (S.H.). E.G. is also supported by Dipartimento di Scienze Biomediche, Università di Sassari, Italy.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.M. (montminy@salk.edu).

A DNA damage checkpoint response in telomere-initiated senescence

Fabrizio d'Adda di Fagagna^{1*}, Philip M. Reaper¹, Lorena Clay-Farrace¹, Heike Fiegler², Philippa Carr², Thomas von Zglinicki³, Gabriele Saretzki³, Nigel P. Carter² & Stephen P. Jackson¹

¹The Wellcome Trust/Cancer Research UK Institute of Cancer and Developmental Biology, University of Cambridge, Cambridge CB2 1QR, UK

²The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

³Henry Wellcome Laboratory of Biogerontology, Institute for Ageing and Health, University of Newcastle upon Tyne, Newcastle NE4 6BE, UK

* Present address: IFOM-FIRC Institute of Molecular Oncology, via Adamello 16, 20139 Milan, Italy

Most human somatic cells can undergo only a limited number of population doublings *in vitro*¹. This exhaustion of proliferative potential, called senescence, can be triggered when telomeres—the ends of linear chromosomes—cannot fulfil their normal protective functions. Here we show that senescent human fibroblasts display molecular markers characteristic of cells bearing DNA double-strand breaks. These markers include nuclear foci of phosphorylated histone H2AX and their co-localization with DNA repair and DNA damage checkpoint factors such as 53BP1,

MDC1 and NBS1. We also show that senescent cells contain activated forms of the DNA damage checkpoint kinases CHK1 and CHK2. Furthermore, by chromatin immunoprecipitation and whole-genome scanning approaches, we show that the chromosome ends of senescent cells directly contribute to the DNA damage response, and that uncapped telomeres directly associate with many, but not all, DNA damage response proteins. Finally, we show that inactivation of DNA damage checkpoint kinases in senescent cells can restore cell-cycle progression into S phase. Thus, we propose that telomere-initiated senescence reflects a DNA damage checkpoint response that is activated with a direct contribution from dysfunctional telomeres.

Senescence in primary human diploid fibroblasts (HDFs) can be triggered by telomere erosion². We speculated that shortened, dysfunctional telomeres might activate events that, under other circumstances, are elicited by DNA double-strand breaks (DSBs). Cells react to DSBs by triggering the DNA damage checkpoint response, which arrests cell-cycle progression until the DNA damage has been removed. Enforcement of the DNA damage checkpoint relies on the concerted activities of the upstream protein kinases ATM and ATR, the downstream transducer kinases CHK1 and CHK2, and mediators such as 53BP1 and MDC1 (refs 3–7). In addition to targeting proteins such as RAD17, SMC1 and p53, ATM and ATR phosphorylate S139 of histone H2AX molecules flanking the site of DNA damage³. Phosphorylated H2AX (γ -H2AX)—a

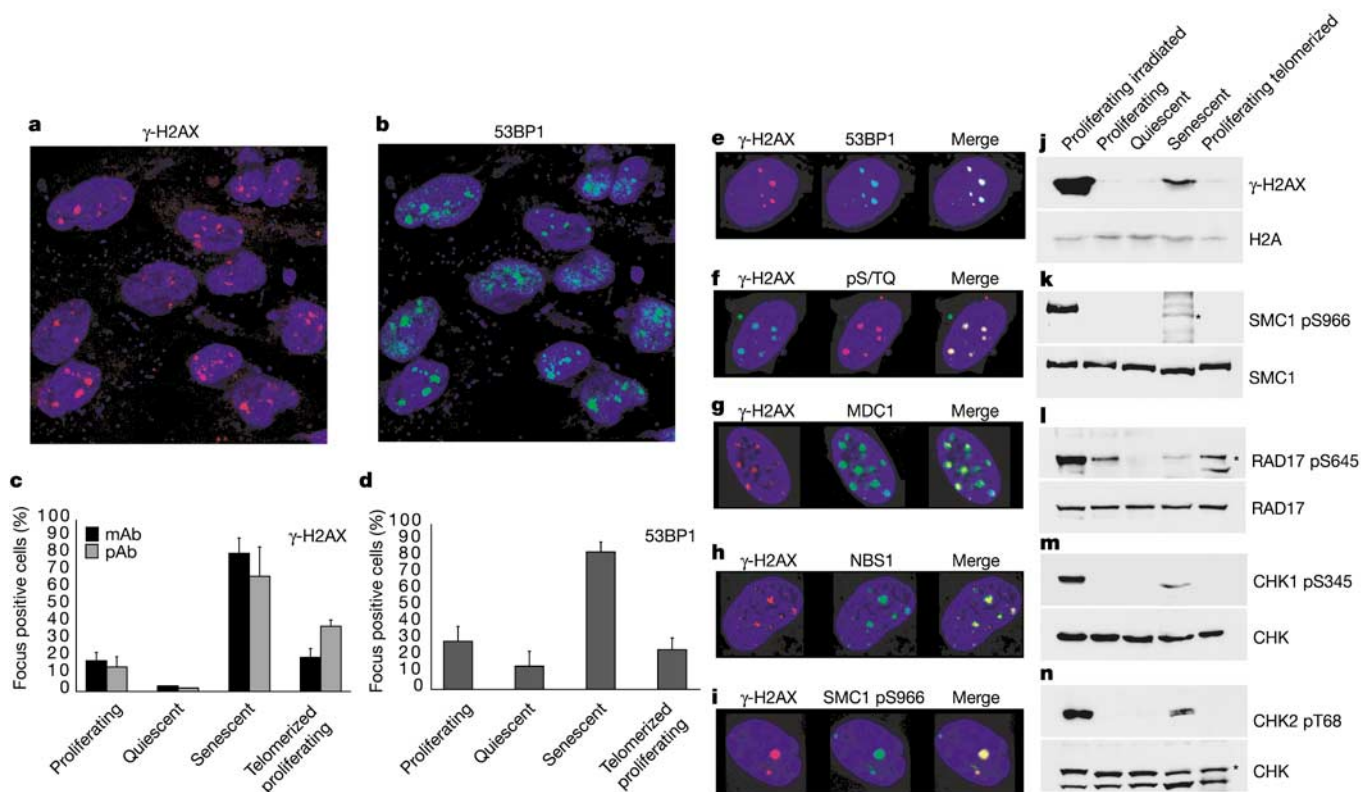


Figure 1 A DNA damage response in senescent cells. **a, b**, γ -H2AX and 53BP1 foci in senescent cells. **c, d**, Quantification of γ -H2AX and 53BP1 foci. γ -H2AX was stained with a mouse monoclonal or an affinity-purified rabbit polyclonal antibody. Histograms indicate the percentage of cells with at least one clearly detectable focus in young (population doublings < 30) proliferating, quiescent, senescent (population doublings about 45) and telomerized MRC5 (population doublings > 70) cell cultures. On average more than 200 cells were screened per point in independent experiments. Error bars represent s.d. **e–i**, Senescent fibroblasts display 53BP1, pS/TQ, MDC1, NBS1 and SMC1 pS966 foci that co-localize with γ -H2AX foci. **j–n**, Whole-cell extracts were prepared from irradiated and untreated proliferating, quiescent, senescent and proliferating telomerized cells.

j, γ -H2AX accumulation in senescent cells. **k**, SMC1 S966 phosphorylation in senescent cells. The higher background and apparent faster mobility of SMC1 in the senescent cell lane is due to the increased extract loading necessary to normalize the amount of SMC1 (Supplementary Fig. 3c). **l**, RAD17 S645 phosphorylation in senescent cells; this marker is also observed in undamaged proliferating and late population-doubling telomerized cells, probably reflecting a role in normal S-phase progression. **m**, CHK1 S345 phosphorylation in senescent cells. Because CHK1 is downregulated in non-proliferating cells (Supplementary Fig. 3c) the amounts of this protein were normalized before analysis. **n**, CHK2 T68 phosphorylation in senescent cells. Each panel derives from immunoblotting of the same membrane. Where shown, an asterisk indicates the specific band.

robust marker for cellular DSBs—then facilitates the focal assembly of checkpoint and DNA repair factors.

To see whether the DNA damage response is triggered during senescence, we used MRC5 and BJ cells; two HDF lines that undergo senescence triggered by telomere shortening (ref. 2 and our own unpublished data). Notably, immunofluorescence experiments revealed that in senescent cultures nearly all cells had clearly detectable γ -H2AX and 53BP1 foci (Fig. 1a, b). Confocal microscopy indicated that these foci of γ -H2AX and 53BP1 were largely coincident (Fig. 1e). Quantitative analyses revealed that only a small proportion of young proliferating or quiescent MRC5 cells had at least one detectable γ -H2AX or 53BP1 focus (Fig. 1c, d), and that the proportion of foci increased progressively as cultures approached senescence (data not shown). Significantly, proliferat-

ing telomerized MRC5 cultures—cells forcibly expressing telomerase and grown in culture for more population doublings than the senescent cells—displayed levels of γ -H2AX and 53BP1 foci similar to those of young cells. Analogous data were obtained with proliferating, quiescent and senescent BJ cells (Supplementary Fig. 1).

Because focus formation by γ -H2AX and 53BP1 after ionizing radiation depends on ATM and ATR³, the above data suggested that these kinases are active in senescent cells. Consistent with this, and in line with work implicating ATM in senescence triggered by telomere shortening or uncapping^{8–11}, senescent cells exhibited focal staining by affinity-purified antibodies against the ATM/ATR phosphorylated consensus target sequence (phosphorylated (p)S/TQ; Fig. 1f). Moreover, γ -H2AX foci in senescent cells co-

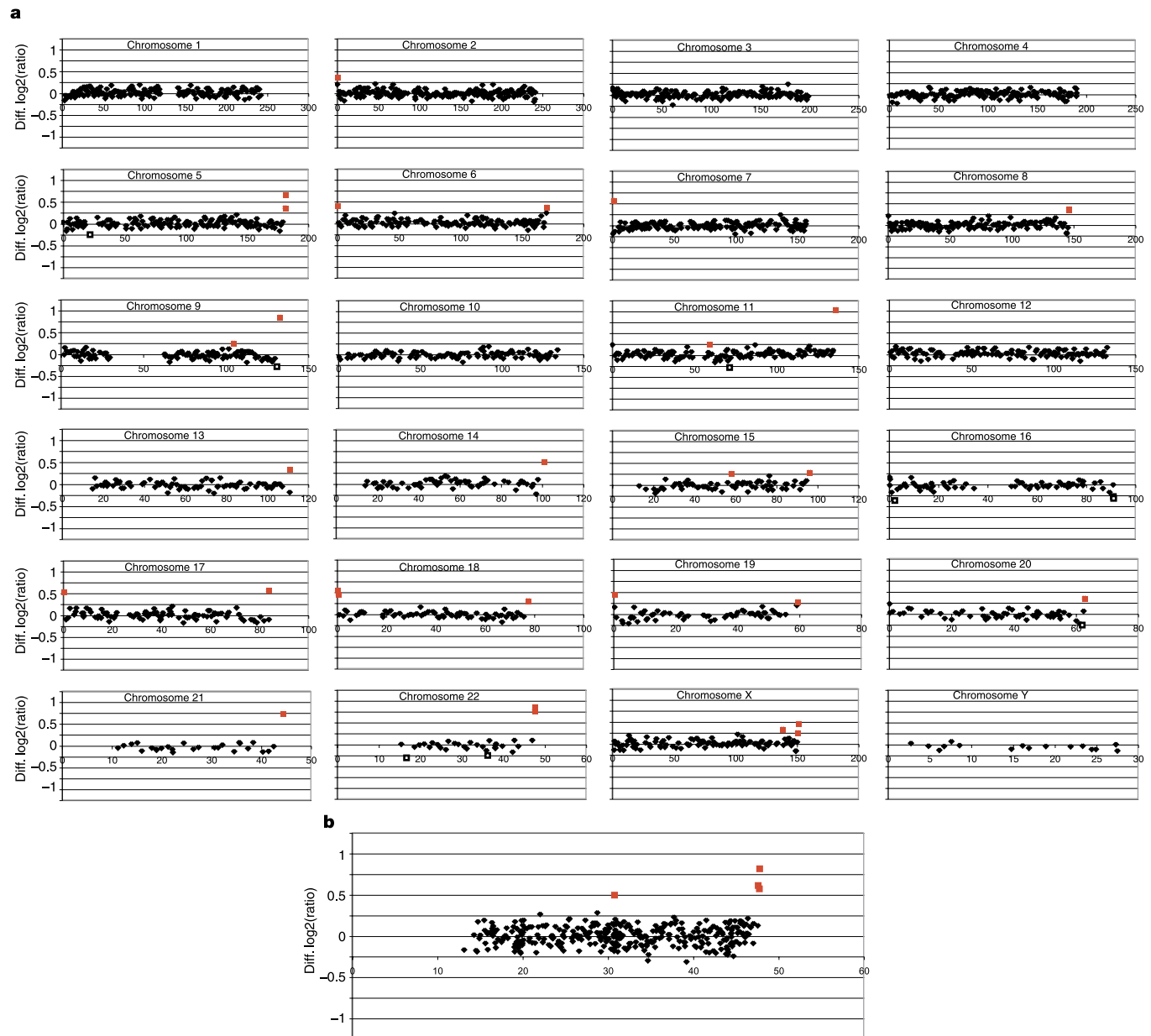


Figure 2 Chromosome ends are directly associated with γ -H2AX in senescent HDFs. Differentially labelled material immunoprecipitated from senescent cells and from quiescent cells was analysed using a whole-genome DNA microarray (a) and an array of overlapping DNA clones for chromosome 22q (b). The y axis represents the difference between the ratios of immunoprecipitated versus input DNA of senescent and quiescent cells along the genome (Diff. log₂(ratio)). The x axis describes the distance in megabases.

Red squares represent clones corresponding to regions enriched in immunoprecipitated material from senescent cells (ratio difference above the 99% confidence interval); black squares represent clones reporting a ratio difference below the 99% confidence interval. We noted a small number of clones (open squares) that fell marginally below the 99% confidence interval, as would be expected on a statistical basis.

localized with foci of the DNA damage checkpoint protein MDC1, the checkpoint and DSB repair factor NBS1, and the S966 phosphorylated (pS966) form of SMC1 (Fig. 1g–i). By contrast, subnuclear compartments known as PML bodies only incompletely overlapped with γ -H2AX foci, despite their increased presence in senescent cells (Supplementary Fig. 2). Taken together, these results reveal that widespread focus formation by γ -H2AX and by a range of DNA DSB response factors is associated with replicative senescence in the two different HDF cell lines tested. It is, therefore, likely to be a general feature of HDFs, and possibly other cell types, undergoing telomere-initiated senescence. We propose the name SDF (for senescence-associated DNA damage focus) for this novel subnuclear structure.

To extend our characterization of the DNA damage response in senescent cells, we used immunoblotting. In addition to confirming the accumulation of γ -H2AX and pS966 SMC1 (Fig. 1j, k), this analysis revealed elevated amounts of p53, its S15 phosphorylated form and p21 in senescent cells (Supplementary Fig. 3a, b). Moreover, we consistently observed S645 phosphorylation (pS645) of RAD17—an event mediated by ATM/ATR and necessary for full checkpoint execution after ionizing radiation³—in senescent cells but not in quiescent cells (Fig. 1l). Thus, in non-dividing cells RAD17 pS645 is specifically associated with the senescent state. Notably, senescent cells but not young proliferating, quiescent or telomerized cells also possessed CHK1 and CHK2 phosphorylated on S345 and T68, respectively (Fig. 1m, n). As phosphorylation of these sites by ATM/ATR is required for full execution of DNA-damage-induced cell-cycle arrest, these results suggest that telomere-initiated senescence triggers the full repertoire of the DNA damage response.

The DNA damage response in senescent cells may arise directly from dysfunctional telomeres or from broken dicentric chromosomes generated after the fusion of two eroded chromosome ends, or from both of these. To ascertain whether dysfunctional telomeres directly engage the DNA damage response in senescent cells, we investigated the chromosomal localization of γ -H2AX. To this end we performed chromatin immunoprecipitation (ChIP) on cross-linked chromatin from senescent MRC5 cells with antibodies against γ -H2AX, then used the immunoprecipitated DNA to probe microarrays containing DNA fragments spanning the entire human genome (at 1 clone per megabase (Mb) resolution). Figure 2 displays the results obtained by this approach in the form of the

difference between the ratios of immunoprecipitated versus input DNA for senescent and quiescent cells. Notably, this analysis revealed that γ -H2AX accumulated at a subset of subtelomeric regions in the senescent cell population, with a bias towards chromosome ends known to bear short telomeres¹². These enrichments were not observed in the absence of γ -H2AX antibodies (data not shown). To map γ -H2AX association at higher resolution, we used a higher density array of overlapping DNA fragments (tiling path) spanning the length of chromosome 22q. This confirmed the association of γ -H2AX with the chromosome 22q telomere and revealed that γ -H2AX spreads for more than 270 kilobases inward from this chromosome end (Fig. 2b). Therefore, despite the heterogeneous and stochastic nature of telomere shortening, these studies establish that telomeres contribute directly to the DNA damage response in senescent fibroblasts.

To complement the above studies and to study the molecular events immediately preceding and concomitant with senescence, we made use of a system in which senescence can be induced synchronously¹³. In this system, the endogenous telomere-capping factor TRF2 is stripped off telomeres by expression of a truncated dominant-negative form of the protein (TRF2 Δ B Δ M; Fig. 3a). The ensuing progressive telomeric uncapping, while preserving normal telomere length, triggers senescence in human fibroblasts as well as apoptosis in other cell types^{10,11}. As was the case for naturally senescing cell populations, we found that TRF2 Δ B Δ M expression led to accumulation of γ -H2AX (Fig. 3b), pS966 SMC1 (Fig. 3c), pS645 RAD17 (Fig. 3d), the activated phosphorylated forms of CHK1 and CHK2 (Fig. 3e, f), and nuclear foci of 53BP1 and γ -H2AX (data not shown).

We next used ChIPs to test for the recruitment of DNA repair and checkpoint factors to telomeric DNA in the TRF2 Δ B Δ M system. Figure 3g illustrates the specific fold increase in protein association with telomeric DNA after TRF2 Δ B Δ M induction (see Supplementary Fig. 4 for representative dot-blot hybridizations). As expected, TRF2 Δ B Δ M expression reduced endogenous TRF2 association with telomeric DNA by nearly 50%, whereas the association of TRF1—a different telomere-binding protein—was not significantly changed, thus providing a control for telomeric DNA integrity. Most notably, TRF2 Δ B Δ M induction led to a progressive and substantial increase in telomeric DNA retrieved by antibodies against γ -H2AX and 53BP1. Furthermore, increased telomeric associations of NBS1 and of the DNA damage checkpoint protein RAD1 (ref. 14) were also

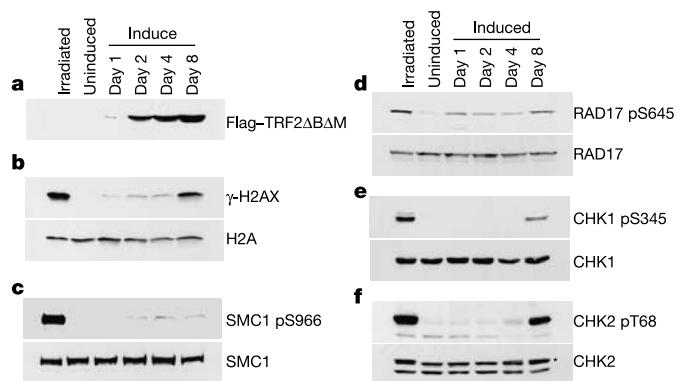
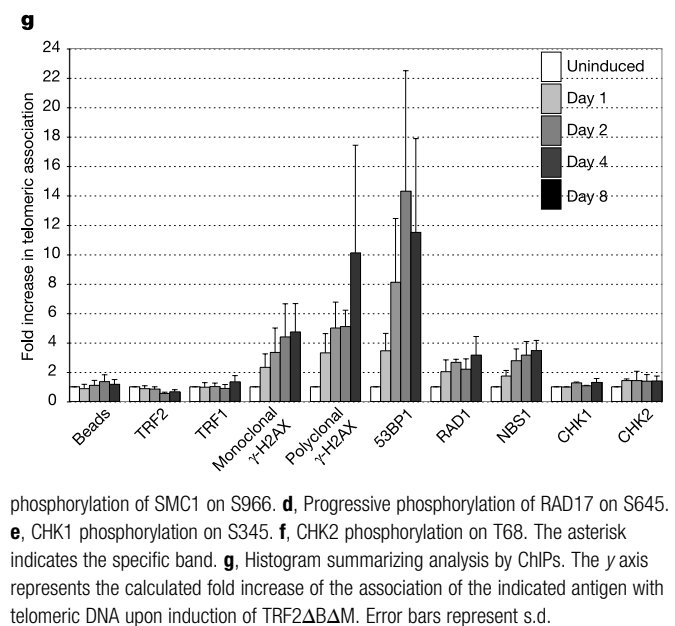


Figure 3 A DNA damage response is generated upon telomere uncapping. Immortalized human fibroblasts (T19) were induced to express TRF2 Δ B Δ M for up to 8 days and whole-cell extracts were generated at the indicated time points. Extracts were also prepared from uninduced cells and uninduced irradiated cells. **a**, Induced expression of Flag-TRF2 Δ B Δ M. **b**, Progressive accumulation of γ -H2AX. **c**, Progressive



phosphorylation of SMC1 on S966. **d**, Progressive phosphorylation of RAD17 on S645. **e**, CHK1 phosphorylation on S345. **f**, CHK2 phosphorylation on T68. The asterisk indicates the specific band. **g**, Histogram summarizing analysis by ChIPs. The y axis represents the calculated fold increase of the association of the indicated antigen with telomeric DNA upon induction of TRF2 Δ B Δ M. Error bars represent s.d.

observed. Significantly, although TRF2 Δ B Δ M induction triggered CHK1 and CHK2 phosphorylation (Fig. 3e, f), no telomere binding by these factors or their active phosphorylated forms was observed (Fig. 3g and data not shown). This is consistent with recent data demonstrating that CHK2 does not stably associate with sites of DNA damage¹⁵. These results establish that a range of DNA damage response proteins associate with dysfunctional telomeres and contribute to DNA damage checkpoint activation.

DNA damage checkpoint responses can be dampened by expression of kinase-dead (KD) forms of ATM, ATR, CHK1 and CHK2 (refs 16–19). To investigate whether maintenance of senescence requires sustained DNA damage signalling, we microinjected senescent BJ cells with combinations of plasmids expressing ATM-KD, ATR-KD, CHK1-KD and CHK2-KD, or with an equal amount of parental vector, in parallel and on the same tissue-culture dish. Cells were then monitored for S-phase entry by assessing 5-bromodeoxyuridine (BrdU) incorporation (Fig. 4). Notably, although <3% of senescent cells microinjected with empty vector incorporated BrdU, expression of all four KD constructs induced BrdU incorporation in 17% of microinjected cells. Furthermore, expression of pairwise combinations of ATM-KD and ATR-KD, or CHK1-KD and CHK2-KD, induced BrdU incorporation in 15% and 12% of microinjected cells, respectively. BrdU-positive cells exhibited reduced levels of γ -H2AX (data not shown) and stained positive for expression of MCM7 (Supplementary Fig. 5)—a highly specific marker for chromosomal DNA replication and cell proliferation that is downregulated in non-proliferating cells²⁰. Similar data were produced by small interfering (si)RNA-mediated knock-down of the same set of checkpoint proteins (Supplementary Fig. 6). These data therefore confirm the involvement of the DNA damage checkpoint apparatus in the senescence programme, and reveal that efficient maintenance of cell-cycle arrest in senescent cells depends on the continued activity of DNA damage checkpoint kinases.

On the basis of the above findings, we propose that telomere-initiated senescence reflects a permanent cell-cycle arrest enforced when the DNA damage checkpoint pathway is activated above a certain threshold by critically short telomeres and probably also by broken dicentric chromosomes. Telomere-initiated senescence can, therefore, be likened to well-characterized responses elicited by ionizing-radiation-induced DNA DSBs. Our ability to detect SDFs more than two months after cells have entered senescence (data not shown) combined with the results of the checkpoint

interference experiments suggest that the DNA damage response is actively maintained in the senescent state. This DNA damage checkpoint model for replicative senescence may explain why ionizing radiation and some other DNA-damaging agents yield a permanent cell-cycle arrest and senescent phenotype in HDFs²¹, and might explain the extended lifespan of HDFs grown in the presence of antioxidants or under hypoxia—a condition that reduces DNA damage by oxygen-derived free radicals. Conversely, it might help to explain why certain DNA-repair-deficient knockout mice and human cell lines senesce prematurely²², and why mice with shortened telomeres are hypersensitive to DNA damage^{23,24}. Finally, we expect that these phenomena occur in animal tissues undergoing replicative senescence *in vivo*. This model may therefore be of relevance in terms of the proposed role of senescence as a tumour-suppressor mechanism²⁵, particularly as several of the checkpoint factors that respond to telomere erosion are known tumour suppressors²⁶. □

Methods

Cell culture

MRC5 and BJ HDFs were grown under standard tissue culture conditions in DMEM supplemented with 10% fetal bovine serum. Senescence was evaluated by failure to reach confluence after three weeks in culture from the last 1:2 passage, failure of >95% cells to incorporate BrdU after a 24-h period, expression of senescence-associated β -galactosidase and morphological features. The TRF2 Δ B Δ M-expressing human cell line T19 (a gift from T. de Lange) was maintained as previously described¹³. Irradiated cells were analysed 1 h after treatment with 20 Gy.

Immunofluorescence microscopy

Cells were fixed either with 2% paraformaldehyde in PBS for 10 min and permeabilized with 0.5% NP-40 for 10 min, or with 50% methanol/50% acetone for 2 min at room temperature. Coverslips were blocked in PBG (0.2% cold water fish gelatine, 0.5% BSA in PBS) and incubated with primary antibody. Cells were then washed and incubated with secondary antibodies (Jackson Laboratories) and DNA dye TOTO-3 (Molecular Probes). Confocal sections were obtained with a Bio-RAD confocal laser microscope by sequential scanning. Comparative immunofluorescence analyses for proliferating, quiescent, senescent and telomerized cells were always carried out in parallel.

Immunoblotting

Extracts were prepared by lysis in TEB150 buffer (50 mM HEPES pH 7.4, 150 mM NaCl, 2 mM MgCl₂, 5 mM EGTA pH 8, 1 mM dithiothreitol, 0.5% Triton X-100, 10% glycerol, 1 mM Na₃VO₄, 1 μ M microcystin-LR, $\times$ 1 Complete Cocktail (Roche)) or SDS–polyacrylamide gel electrophoresis (PAGE) sample buffer. Proteins were resolved by SDS–PAGE, transferred to nitrocellulose and probed. Results shown are representative of at least three independent experiments.

Chromatin immunoprecipitation

In vivo cross-linking, chromatin purification and immunoprecipitations were as previously described²⁷. For Fig. 3g, ratios between telomeric DNA immunoprecipitated from induced and uninduced cells were calculated. After normalizing resulting data to the corresponding ratio of an Alu probe (a measure of nonspecific DNA binding), the fold increase in association of proteins with telomeric DNA was calculated. Experiments were performed at least three times per antigen.

DNA microarray preparation

Microarrays were constructed as described previously²⁸. For the whole-genome array, large insert clones were selected from the Golden Path—the minimum set of overlapping clones used to generate the sequence of the human genome—to be spaced at approximately 1-Mb intervals. For the chromosome 22 tiling path array, 447 clones were selected, 444 of which were chosen from the Golden Path; the other three clones were identified by their BAC end sequence position. Clone details for the arrays can be obtained from http://www.ensembl.org/Homo_sapiens/cytoview/. To construct arrays, clone DNA was isolated, amplified by DOP-PCR and spotted in either duplicate (1-Mb array) or triplicate (tiling path) onto 3D-link activated slides (Motorola) using a MicroGrid II arrayer (BioRobotics).

DNA labelling, hybridization and analysis

DNA labelling and hybridization were essentially as described previously²⁸. Arrays were scanned with an Agilent scanner (Agilent Technologies) and fluorescent intensities extracted using SPOT software. After local background subtraction, signal intensities were normalized to a 1:1 ratio (immunoprecipitated to input DNA) by dividing the ratio of every data point by the mean ratio of all autosomal clones (1-Mb array) or all clones (tiling path). All hybridizations were performed as dye-label reversal experiments, and the mean ratio (immunoprecipitated:input DNA) of both hybridizations was calculated. The difference between the ratios generated with material immunoprecipitated from senescent cells and quiescent cells was then plotted against the clone position. Ratios were considered as significant if they fell outside the 99% confidence intervals.

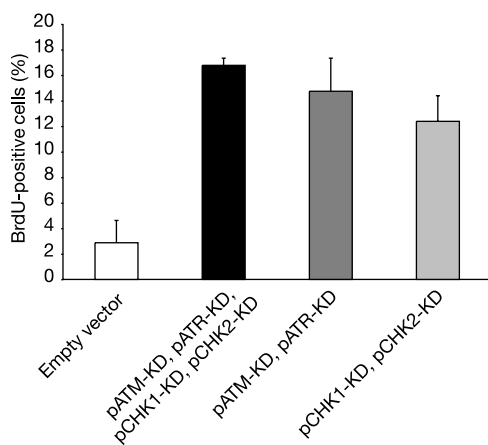


Figure 4 Checkpoint inactivation in senescent cells induces S-phase progression. The y axis shows the percentage of BrdU-positive senescent BJ cells microinjected with plasmids expressing dominant-negative alleles of the indicated DNA damage checkpoint kinases or empty vector. Values reported are the results of at least three independent experiments using two different batches of senescent cells. A total of 850 microinjected cells were counted. Error bars represent s.d.

Antibodies

We used mouse and rabbit anti- γ -H2AX, anti-CHK1 (Upstate Biotechnology); anti-pS/TQ, anti-CHK1 pS345, anti-CHK2 pT68, anti-p53 pS15 (Cell Signalling Technology); anti-NBS1, anti-CHK2 (Oncogene); anti-SMC1 pS966, anti-SMC1 (Bethyl Laboratories); anti-PML, anti-RAD17 (H-300), anti-H2A (H-124), anti-CHK1 (G-4), anti-p21 (C-19), anti-TRF1 (C-19), anti-TRF2 (N-20) (Santa Cruz Biotechnology); anti-Flag (M2) (Sigma); and anti-CHK2 (Abcam) antibodies. Anti-MDC1, anti-NBS1 (ref. 7); anti-RAD1 (ref. 14); anti-MCM7, anti-53BP1, anti-RAD17 pS645; and anti-p53 (DO-1) antibodies were gifts from R. Laskey, T. Halazonetis, R. Abraham and T. Kouzarides, respectively.

Plasmids

pATM-KD, pATR-KD and pCHK1-KD were gifts from M. Kastan, R. Abraham and J. Bartek, respectively. pCHK2-KD, a gift from B. Williams, was generated by PCR amplification of the CHK2 open reading frame, a gift from S. Elledge, followed by cloning into a mammalian expression vector and site-directed mutagenesis. pH2B-YFP, a gift from C. Aquaviva, expresses histone H2B fused to yellow fluorescent protein.

Microinjection

Cells were plated on poly-L-lysine-coated glass dishes (Biopetechs) and incubated in Nut Mix F-12 (Ham) medium (Gibco) plus 10% newborn calf serum. Each plasmid and pH2B-YFP as an injection marker was microinjected at a concentration of $100 \text{ ng } \mu\text{l}^{-1}$ into senescent cell nuclei using Fentotips (Eppendorf). BrdU incorporation over a 3.5-day period was monitored by Amersham Cell Proliferation Kit.

siRNA

siRNAs were synthesized by Dharmacon and were transfected by Oligofectamine (Invitrogen). Oligonucleotide sequences were as in refs 7, 29, 30.

Received 1 May; accepted 13 October 2003; doi:10.1038/nature02118.

Published online 5 November 2003.

- Hayflick, L. & Moorhead, P. S. The limited *in vitro* lifetime of human diploid cell strains. *Exp. Cell Res.* **25**, 585–621 (1961).
- Bodnar, A. G. *et al.* Extension of life-span by introduction of telomerase into normal human cells. *Science* **279**, 349–352 (1998).
- Shiloh, Y. ATM and related protein kinases: safeguarding genome integrity. *Nature Rev. Cancer* **3**, 155–168 (2003).
- Nyberg, K. A., Michelson, R. J., Putnam, C. W. & Weinert, T. A. Toward maintaining the genome: DNA damage and replication checkpoints. *Annu. Rev. Genet.* **36**, 617–656 (2002).
- Lou, Z., Minter-Dykhouse, K., Wu, X. & Chen, J. MDC1 is coupled to activated CHK2 in mammalian DNA damage response pathways. *Nature* **421**, 957–961 (2003).
- Stewart, G. S., Wang, B., Bignell, C. R., Taylor, A. M. R. & Elledge, S. J. MDC1 is a mediator of the mammalian DNA damage checkpoint. *Nature* **421**, 961–966 (2003).
- Goldberg, M. *et al.* MDC1 is required for the intra-S-phase DNA damage checkpoint. *Nature* **421**, 952–956 (2003).
- Webley, K. *et al.* Posttranslational modifications of p53 in replicative senescence overlapping but distinct from those induced by DNA damage. *Mol. Cell. Biol.* **20**, 2803–2808 (2000).
- de Lange, T. Protection of mammalian telomeres. *Oncogene* **21**, 532–540 (2002).
- Karlseder, J., Broccoli, D., Dai, Y., Hardy, S. & de Lange, T. p53- and ATM-dependent apoptosis induced by telomeres lacking TRF2. *Science* **283**, 1321–1325 (1999).
- Smogorzewska, A. & de Lange, T. Different telomere damage signaling pathways in human and mouse cells. *EMBO J.* **21**, 4338–4348 (2002).
- Martens, U. M. *et al.* Short telomeres on human chromosome 17p. *Nature Genet.* **18**, 76–80 (1998).
- van Steensel, B., Smogorzewska, A. & de Lange, T. TRF2 protects human telomeres from end-to-end fusions. *Cell* **92**, 401–413 (1998).
- Freire, R. *et al.* Human and mouse homologs of *Schizosaccharomyces pombe* rad1⁺ and *Saccharomyces cerevisiae* RAD17: linkage to checkpoint control and mammalian meiosis. *Genes Dev.* **12**, 2560–2573 (1998).
- Lukas, C., Falck, J., Bartkova, J., Bartek, J. & Lukas, J. Distinct spatiotemporal dynamics of mammalian checkpoint regulators induced by DNA damage. *Nature Cell Biol.* **5**, 255–260 (2003).
- Canman, C. E. *et al.* Activation of the ATM kinase by ionizing radiation and phosphorylation of p53. *Science* **281**, 1677–1679 (1998).
- Cliby, W. A. *et al.* Overexpression of a kinase-inactive ATR protein causes sensitivity to DNA-damaging agents and defects in cell cycle checkpoints. *EMBO J.* **17**, 159–169 (1998).
- Heffernan, T. P. *et al.* An ATR- and Chk1-dependent S checkpoint inhibits replicon initiation following UVC-induced DNA damage. *Mol. Cell. Biol.* **22**, 8552–8561 (2002).
- Chehab, N. H., Malikzay, A., Appel, M. & Halazonetis, T. D. Chk2/hCds1 functions as a DNA damage checkpoint in G(1) by stabilizing p53. *Genes Dev.* **14**, 278–288 (2000).
- Stoeber, K. *et al.* DNA replication licensing and human cell proliferation. *J. Cell Sci.* **114**, 2027–2041 (2001).
- Schmitt, C. A. Senescence, apoptosis and therapy—cutting the lifelines of cancer. *Nature Rev. Cancer* **3**, 286–295 (2003).
- Parrinello, S. *et al.* Oxygen sensitivity severely limits the replicative lifespan of murine fibroblasts. *Nature Cell Biol.* **5**, 741–747 (2003).
- Wong, K. K. *et al.* Telomere dysfunction impairs DNA repair and enhances sensitivity to ionizing radiation. *Nature Genet.* **26**, 85–88 (2000).
- Goytisolo, F. A. *et al.* Short telomeres result in organismal hypersensitivity to ionizing radiation in mammals. *J. Exp. Med.* **192**, 1625–1636 (2000).
- Campisi, J. Cellular senescence as a tumor-suppressor mechanism. *Trends Cell Biol.* **11**, S27–S31 (2001).
- Hoeijmakers, J. H. Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366–374 (2001).
- d'Adda di Fagnola, F. *et al.* Effects of DNA nonhomologous end-joining factors on telomere length and chromosomal stability in mammalian cells. *Curr. Biol.* **11**, 1192–1196 (2001).

- Fiegler, H. *et al.* DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosom. Cancer* **36**, 361–374 (2003).
- Zhao, H., Watkins, J. L. & Piwnicka-Worms, H. Disruption of the checkpoint kinase 1/cell division cycle 25A pathway abrogates ionizing radiation-induced S and G2 checkpoints. *Proc. Natl Acad. Sci. USA* **99**, 14795–14800 (2002).
- Sorensen, C. S. *et al.* Chk1 regulates the S phase checkpoint by coupling the physiological turnover and ionizing radiation-induced accelerated proteolysis of Cdc25A. *Cancer Cell* **3**, 247–258 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. Abraham, C. Aquaviva, J. Bartek, R. Laskey, N. La Thangue, T. de Lange, S. Elledge, T. Halazonetis, M. Kastan and T. Kouzarides for sharing reagents; K. Woodfine for constructing the chromosome 22 array; M. Goldberg for help with immunofluorescence; B. Williams for generating pCHK2-KD; J. Bradbury and K. Dry for editorial help; and D. Baird, J. Bartek, M. Foiani, J. Pines and all the S.P.J. laboratory for suggestions. Ed'A.d.F. was supported by a Cancer Research UK grant, and P.M.R. is supported by a studentship from Cancer Research UK. T.v.Z. acknowledges support from Research into Ageing UK.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.P.J. (spj13@mole.bio.cam.ac.uk) or F.d'A.d.F. (dadda@ifom-firc.it).

DNA self-recognition in the structure of Pot1 bound to telomeric single-stranded DNA

Ming Lei¹, Elaine R. Podell¹, Peter Baumann² & Thomas R. Cech¹

¹Howard Hughes Medical Institute, Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado 80309-0215, USA

²Stowers Institute for Medical Research, 1000 East 50th Street, Kansas City, Missouri 64110, USA

Telomeres, specialized protein–DNA complexes that cap the ends of linear chromosomes, are essential for protecting chromosomes from degradation and end-to-end fusions^{1,2}. The Pot1 (protection of telomeres 1) protein is a widely distributed eukaryotic end-capping protein, having been identified in fission yeast, microsporidia, plants and animals^{3,4}. *Schizosaccharomyces pombe* Pot1p is essential for telomere maintenance³, and human POT1 has been implicated in telomerase regulation^{5,6}. Pot1 binds telomeric single-stranded DNA (ssDNA) with exceptionally high sequence specificity⁷, the molecular basis of which has been unknown. Here we describe the 1.9-Å-resolution crystal structure of the amino-terminal DNA-binding domain of *S. pombe* Pot1p complexed with ssDNA. The protein adopts an oligonucleotide/oligosaccharide-binding (OB) fold⁸ with two loops that protrude to form a clamp for ssDNA binding. The structure explains the sequence specificity of binding: in the context of the Pot1 protein, DNA self-recognition involving base-stacking and unusual G–T base pairs compacts the DNA. Any sequence change disrupts the ability of the DNA to form this structure, preventing it from contacting the array of protein hydrogen-bonding groups. The structure also explains how Pot1p avoids binding the vast excess of RNA in the nucleus.

Telomeric DNA usually consists of tandem repeats of a short 5–9-base-pair sequence, varying in length from species to species, and ends with a single-stranded 3' overhang of the G–T-rich strand¹. The ssDNA serves as the substrate for the end-replicating enzyme telomerase⁹. Proteins specific for this telomeric ssDNA have been identified in several organisms and include the telomere end-binding protein (TEBP) of the ciliated protozoan *Oxytricha*

nova^{10,11}, Cdc13p from the budding yeast *Saccharomyces cerevisiae*^{12–14}, and more recently the Pot1 proteins from fission yeast, plant, mouse and human^{3,4}.

The N-terminal domain of *S. pombe* Pot1p, Pot1pN (amino acids 1–185), binds G-rich telomeric ssDNA with the same specificity and even higher affinity than the full-length protein, suggesting that it represents the entire DNA-binding domain^{3,7}. Pot1pN has exceptionally high sequence specificity for the *S. pombe* telomeric sequence GGTTAC⁷. To understand the molecular mechanism of this binding specificity, Pot1pN was crystallized in three different crystal forms in a complex with either the pentanucleotide GGTTA or the hexanucleotide GGTTAC (Supplementary Table 1). The Pot1pN–GGTTA structure of crystal form B was solved using single-wavelength anomalous dispersion (SAD) and refined to a resolution of 1.9 Å ($R_{\text{cryst}} = 22.3\%$, $R_{\text{free}} = 24.7\%$). Electron density was observed for the entire DNA and all except the first four and last ten residues of the protein. The other two crystal forms were then solved by molecular replacement (MR) using a searching model without the ssDNA. Although crystallized under different solution conditions, in different space groups and having different lattice-packing environments, the protein and ssDNA structures were essentially identical in all three crystal forms. The highest-resolution structure of crystal form B will be the focus here, with the information about nucleotide C6 derived from the structure of crystal form C (2.4 Å resolution; Supplementary Fig. 1).

The crystal structure shows that Pot1pN is indeed made up of a compact single domain, the OB fold, consisting of a highly curved

five-stranded anti-parallel β -barrel, as expected from previous sequence-alignment analysis^{3,4}. The ssDNA binds in a basic concave groove, characteristic of OB-fold proteins, formed by one side of the β -barrel and two protruding loops, L12 and L45 (Fig. 1a). The ssDNA runs 5' to 3' from left to right across the cleft as oriented in Fig. 1b; this polarity is identical to that observed in most other OB-fold protein–nucleic acid complexes¹⁵.

In sharp contrast to the rather predictable structure of the protein moiety, the telomeric DNA in the complex adopts a highly compact folded conformation that is entirely unanticipated (Fig. 2). DNA compaction involves both base stacking (Fig. 1c) and two unusual G–T base pairs. There are three sets of pair-wise stacking interactions: G1 stacks with G2, T3 with T4, and A5 with C6; the distance between each set of aromatic rings ranges from 3.5 to 4.2 Å. Each stacking interaction is extended into the protein moiety by hydrophobic or aromatic protein side chains (Leu122, Phe88 and Tyr115, respectively). The T4 deoxyribose ring acts as a platform on which A5, C6 and Tyr115 are stacked. Two non-Watson–Crick base pairs, of a type not previously observed in protein–ssDNA complex structures, involve G1 and G2 making hydrogen bonds between their 2-amino groups and the 4-carbonyl groups of T3 and T4, respectively (Fig. 2b, c). These base-pairing interactions link the adjacent bases in the direction perpendicular to the stacking interactions of G1–G2 and T3–T4, allowing nucleotides G1–T4 to adopt an unusual folded secondary structure that interacts with the protein in its entirety. As a result, the contributions to sequence specificity of the individual nucleotides G1–T4 are no

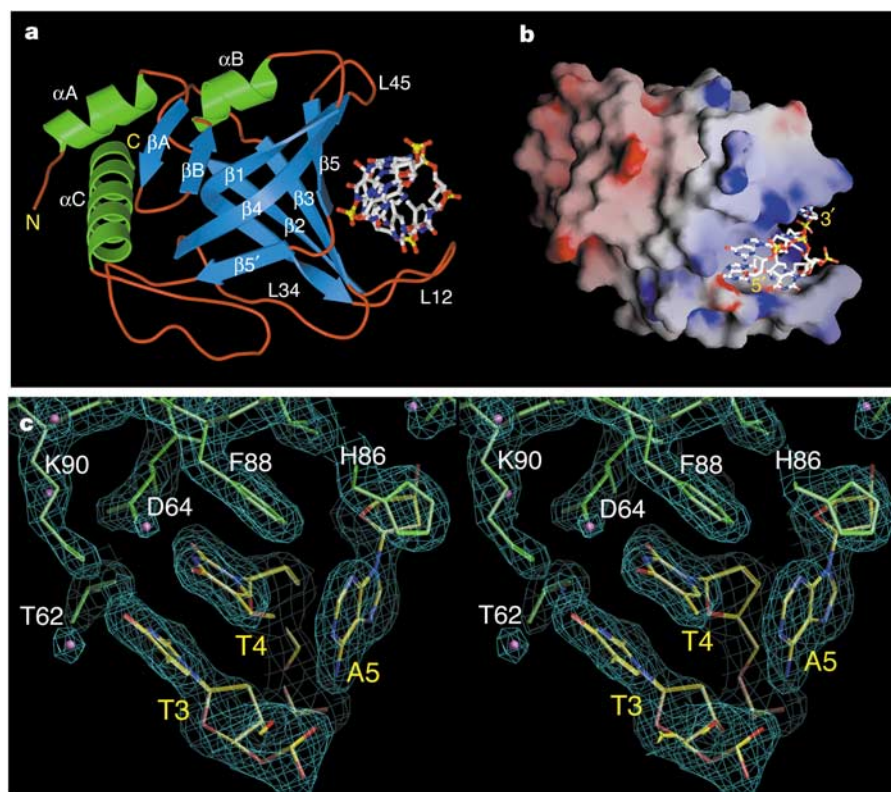


Figure 1 Structure of the Pot1pN–ssDNA complex. **a**, Ribbon diagram with Pot1pN coloured blue (β -strands), green (α -helices) and red (loops). The DNA is shown as a stick representation with carbon coloured white, nitrogen blue, oxygen red and phosphorus yellow. Secondary structure elements of the protein are labelled. This panel was generated using Molscript and Raster 3D (refs 28, 29). **b**, Electrostatic potential surface representation of Pot1pN, calculated with the DNA removed using GRASP³⁰. **c**, Electron

density in the 1.9 Å $2F_o - F_c$ map that shows the stacking interaction of nucleotides T3 and T4 with F88. The refined model of the protein and the DNA is superimposed on the map. Contours are drawn at the 1.5 σ level. Protein is shown in green and the DNA is coloured as in panel **a** except that carbon is yellow. Purple dots (for example, near T62 and D64) are water molecules. This panel was generated using O (ref. 27).

longer independent, and this correlation virtually eliminates the possibility that other sequences might bind to Pot1pN with high affinity.

Yet another type of self-recognition involves hydrogen bonds formed between the base of A5 and the phosphodiester groups of T3 and T4 (Fig. 2d). These hydrogen-bonding interactions appear to stabilize the bent backbone of A5. Indeed, substitution of A5 by T

reduces binding by about 420-fold ($>3 \text{ kcal mol}^{-1}$)⁷. The highly compact, folded conformation of the ssDNA does not require any significantly unusual dihedral angles; the bases are all in *anti* conformations, and sugar puckers are C-2' *endo* as is typical for DNA.

The folded DNA structure is bound by Pot1pN through two types of interaction (Fig. 2e). The protein side chains that provide

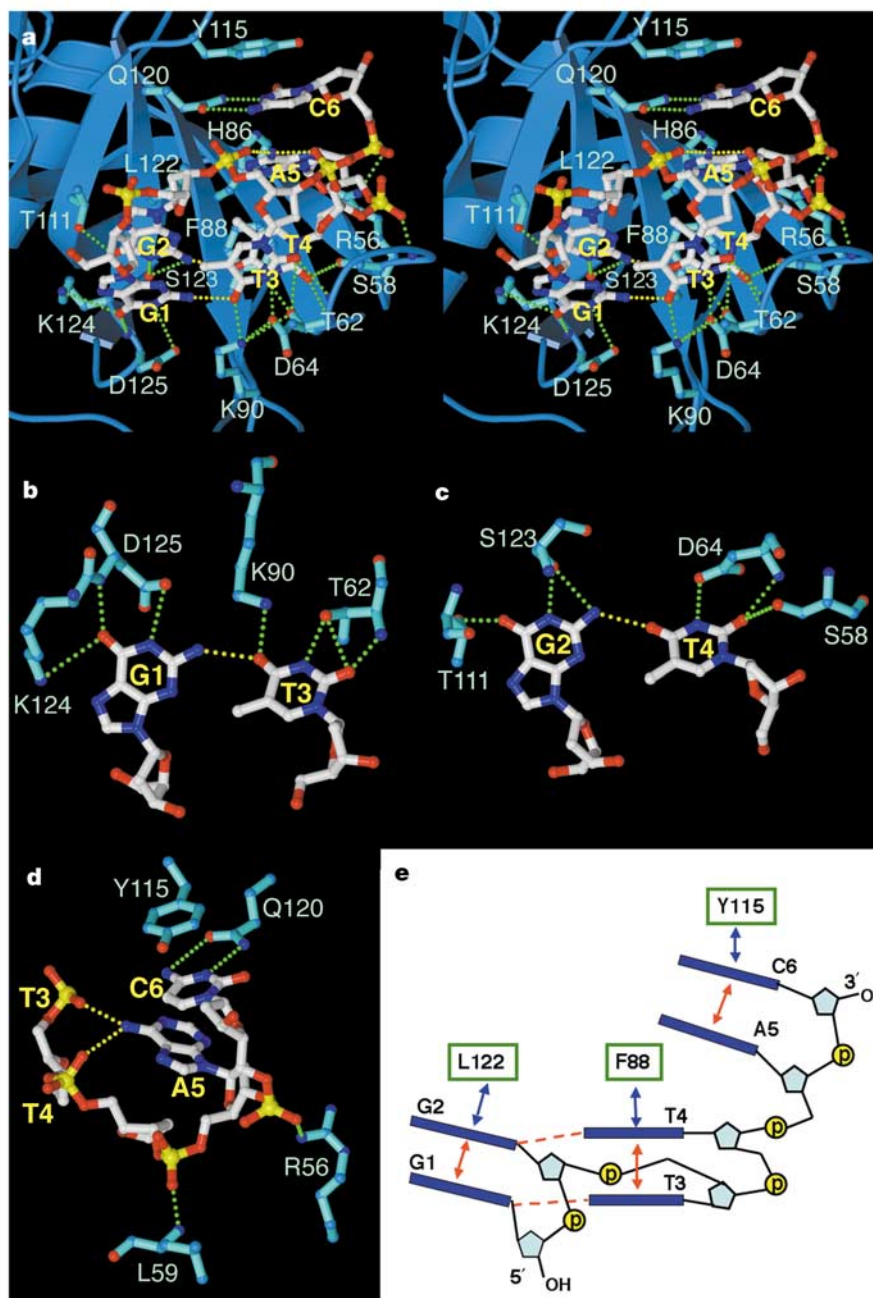


Figure 2 Protein-ssDNA and ssDNA-ssDNA interactions in the Pot1pN-GGTTAC complex. Pot1pN is shown in blue, and the ssDNA is coloured as in Fig. 1a. Protein-DNA intermolecular hydrogen bonds are shown as dotted green lines, and ssDNA intramolecular hydrogen bonds as dotted yellow lines. All panels except **e** were generated using Molscript and Raster 3D (refs 28, 29). **a**, Stereoimage of the protein-ssDNA interactions. The ssDNA is bound in a compact and folded conformation. **b**, **c**, ssDNA self-recognition by G-T base-pairing interactions. The base pairs are oriented such that all Watson-Crick donor/acceptor groups of the bases face the inner side of the binding groove and make extensive hydrogen-bonding interactions with the protein.

d, Interactions of the 3' end of the ssDNA. The hydrogen bonds between A5 and the backbone phosphodiester groups of T3 and T4 represent the second form of self-recognition. **e**, Schematic representation of the Pot1pN-ssDNA interactions. Bases of the DNA are shown as purple bars; phosphodiester groups as yellow circles; sugar rings as cyan pentamers; and protein residues as green boxes. Stacking and van der Waals interactions between bases and protein are shown as blue arrows, and stacking between bases as red arrows. Base-pairing interactions of the ssDNA are shown as red dashed lines.

Table 1 DNA sequence specificity and RNA binding by Pot1pN

Oligonucleotide sequence	K_D (nM)	K_D/K_D^0
A ₇ GGTTAC	480	1
A ₇ IGTTAC	680	1.4
A ₇ GTTTAC	~25,000*	~50
A ₇ GGCTAC	>100,000†	>200
A ₇ GGTCAC	>100,000†	>200
A ₇ GGTTAC	570	1.2
A ₇ GrTTAC	640	1.3
A ₇ GGrUTAC	>100,000†	>200
A ₇ GGTrUAC	>100,000†	>200
A ₇ GGTTAC	760	1.6
A ₇ GGTTAC	710	1.5
A ₇ GGrTTAC	>100,000†	>200
A ₇ GGdUTAC	380	0.7
A ₇ GGTrTAC	3,700	7.7
A ₇ GGTdUAC	~48,000*	~100

Two independent nitrocellulose filter binding assays were performed in 15 mM NaCl, 1 mM DTT, 25 mM Tris-HCl, pH 8.0 at room temperature. The average error of the assay determined from the two measurements was $\pm 9\%$. K_D^0 represents the equilibrium dissociation constant of wild-type telomeric sequence.

*Because incomplete binding was observed at 10 μ M protein, the K_D is only approximate.

†Because no binding was observed at 10 μ M protein, the maximum protein concentration used in the assay, the K_D is estimated to be $>100 \mu$ M.

stacking or van der Waals interactions above each pair of bases presumably contribute sequence-nonspecific binding. In addition, the bases of G1–T4 are bound through an extensive network of 13 hydrogen bonds with protein residues located on the surface of the groove; these seem to be key elements of ssDNA sequence specificity. These hydrogen-bonding interactions are evenly distributed in two layers (with G1 and T3 involved in one, and G2 and T4 in the other), and are entirely mediated by the Watson–Crick donor/acceptor groups of the G and T nucleotides (Fig. 2b, c). Nucleotides A5 and C6 form the other primary area of protein–ssDNA interaction in the complex. With these nucleotides, unlike G1–T4, the phosphodiester backbone of the ssDNA makes important contacts with the protein. The phosphodiester group of A5 forms a hydrogen bond to the main chain amino group of Leu59, whereas that of C6 forms a salt bridge with the amino group of Arg56 (Fig. 2d). Together these interactions appear to bend the backbone of the ssDNA almost 90° away from its initial direction.

In the structure, the 3' end of the ssDNA protrudes from the protein (Fig. 1a, b), in contrast to *O. nova* TEBP, which buries the 3'

end¹¹. However, telomerase requires more than a 3' end for activity, because its RNA subunit normally pairs with the last few bases of the DNA primer. Thus, both Pot1p and TEBP would seem to inhibit telomerase extension in the absence of some sort of activation event.

To test the energetic contributions of the unusual compact ssDNA conformation and the novel G–T pairing, we measured equilibrium binding constants (K_D) for variant oligonucleotides. The crystal structure suggested that substitution of inosine (I), which lacks the 2-amino group of G, should allow protein–DNA interactions despite removing the G–T hydrogen bond. The binding data (Table 1, lines 2 and 3) implied that base pairing of G2–T4 contributes more to formation of the Pot1pN–ssDNA complex than does that of G1–T3. In a second set of experiments, C was substituted for T in position 3 or 4 of the ssDNA, which should disrupt its compact folding by introducing an unfavorable electrostatic clash between the 2-amino group of G1 or G2 and the 4-amino group of C3 or C4. Consistent with this expectation, both T3C and T4C substitutions disrupted the binding between Pot1pN and ssDNA (Table 1). Together, these and previous⁷ binding assays support the energetic importance of the tightly folded DNA conformation observed in the crystal structure.

Pot1p binds ssDNA but not RNA molecules³. How Pot1p differentiates between ssDNA and RNA is an important biological issue, because the amount of RNA in the nucleus vastly exceeds that of ssDNA; r(GGUUAC) RNA sequences would sequester all the Pot1p if its binding were not highly specific for DNA. Binding assays using a series of oligonucleotides with single ribonucleotide substitutions revealed that the specificity for DNA over RNA is not uniform but is conferred by positions 3 and 4 (Table 1). The exclusion of RNA at position 3 is explained by the addition of the hydroxyl group to the sugar ring, whereas the added 2'-hydroxyl group and more importantly the absence of the 5-methyl group contribute at position 4 (Table 1). The crystal structure provides an explanation for these observations. The C-2' atoms of the T3 and T4 deoxyribose rings are only 2.9 and 3.3 Å away from a phosphate oxygen of T3 and the hydroxyl group oxygen of Ser123, respectively, and thus cannot accommodate additional hydroxyl groups. The methyl group of T4 fits in a hydrophobic depression formed by both the protein and the DNA, and its absence in uracil is expected to produce an energetically unfavorable cavity. Thus, the crystal structure and binding experiments show that the specific ssDNA conformation imposed by binding to Pot1pN effectively prevents RNA from binding.

We made site-specific mutations in amino acids directly involved in ssDNA binding or, as a control, in amino acids outside the DNA-

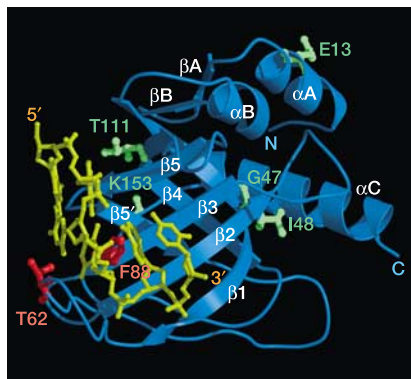


Figure 3 Mutational analysis of residues important for Pot1pN–ssDNA interaction. The protein is shown in blue and the DNA in yellow. The mutated residues are shown in a ball-and-stick model. Red, non-functional mutations T62V (0/84) and F88A (0/156); green, functional mutations E13A (15/24), G47A (88/160), I48T (62/148), T111V (24/48) and K153A (17/36). The numbers in parentheses indicate the results of the complementation analysis (number of isolates that lack wild-type *pot1*⁺ versus total number of isolates tested).

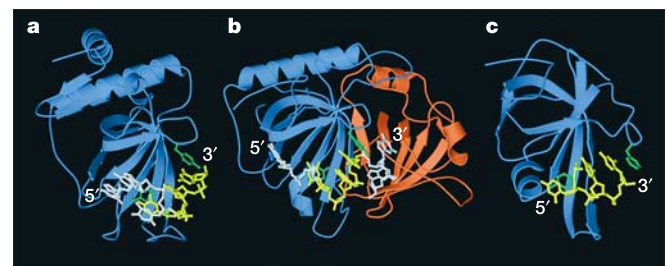


Figure 4 The same OB-fold structure adapted for sequence-specific binding. Comparison of Pot1pN (a), *O. nova* TEBP- α ^{22,23} (b) and human RPA-A^{16,18,19} (c), all bound to ssDNA. The proteins are coloured in blue except for the second OB fold of TEBP- α , which is coloured in red; the two aromatic residues that define the nonspecific motif are coloured in green. The conserved regions of the ssDNA that contact the nonspecific motif are coloured in yellow; the rest are coloured in white. This figure was generated using Molscript and Raster 3D.

binding site. These were cloned into an *S. pombe* expression vector and introduced into a heterozygous diploid strain (*pot1*^{+/pot1}⁻), which was then sporulated and grown to select for plasmid-containing cells. In the case of plasmids encoding wild-type Pot1 or control mutants, the haploids were 50% *pot1*⁺ and 50% *pot1*⁻, indicating that plasmid-expressed Pot1p could complement the *pot1*⁻ deficiency. The sites of tolerated amino-acid substitutions are highlighted in green in Fig. 3. In two cases where mutations were designed to interfere with DNA binding, only cells containing the genomic copy of *pot1*⁺ could be isolated; thus, these mutants (red in Fig. 3) are either non-complementing or confer a decided selective disadvantage. Mutant T111V, designed to eliminate a hydrogen bond to G2, did complement, indicating that loss of this interaction was not essential for function. Strains containing *pot1* point mutants in addition to wild-type genomic *pot1*⁺ showed normal telomere lengths, and all mutant Pot1 proteins were expressed (Supplementary Fig. 2).

Comparison of the three-dimensional structures of Pot1pN (Fig. 4a) and TEBP- α (Fig. 4b) with nonspecific ssDNA-binding proteins, such as human replication protein A (RPA) 70 (ref. 16) and *E. coli* ssDNA-binding protein (SSB)¹⁷, reveals how the same OB-fold structure can be adapted for sequence-specific binding. In RPA (Fig. 4c), proceeding from 5' to 3', a base (yellow) stacks on an aromatic amino acid (green), and the next two bases in the chain then stack on each other and on a second aromatic amino acid (also in green)^{16,18,19}. Both Pot1pN and TEBP- α have this same motif (Fig. 4a, b) with additional nucleotides (white) involved in other interactions. The aromatic amino acids (Phe, Tyr or Trp) are located near the carboxy termini of β 3 and β 4, and lie within the groove on the surface of the OB fold (Supplementary Fig. 3)^{18,20}. For the various SSBs and RPA, these provide the sequence-nonspecific interactions necessary for the proteins to bind to diverse ssDNA sequences during replication. For the sequence-specific telomere-binding proteins, biochemical studies have shown that most of the nucleotides involved in the stacking interactions (A5 and C6 for Pot1pN; T4, G5 and G6 for TEBP- α) make relatively small contributions to sequence specificity^{7,21}. Thus, it appears that the nucleotide-binding motif defined by these aromatic residues is a nonspecific motif for OB-fold ssDNA-binding proteins, and must be supplemented by other interactions in those cases where sequence specificity is required.

In the case of the ciliate TEBP- α , sequence specificity is provided mainly by a second OB fold (coloured red in Fig. 4b) that interacts with nucleotides G7 and G8 (refs 21–23). By contrast, Pot1pN provides a situation where unusual DNA self-recognition is combined with more standard protein–DNA interactions to achieve extraordinary specificity ($\Delta\Delta G^\circ$ more than 4 kcal mol⁻¹ each for substitutions at nucleotides G2, T3 and T4; ref. 7 and Table 1). Given that several OB-fold protein–ssDNA structures are known^{8,15,24}, the finding of a new mode of ssDNA recognition in the Pot1 structure emphasizes that the OB fold is indeed a remarkably versatile framework for protein–ssDNA interactions.

The Pot1 family of proteins seems to provide the most general solution to protecting chromosome ends in eukaryotes⁴. Thus, the structure provided here will serve as a framework for understanding telomere functions at the molecular level. These functions include an essential role for Pot1p in chromosome end-capping in *S. pombe*³, its contribution to the regulation of telomerase in human cells^{5,6}, and its proposed role in switching telomeres between the protective t-loop structure² and an open structure accessible to telomerase⁷. □

Methods

Sample preparation and crystallization

Pot1pN was overexpressed and purified from *E. coli*, and DNA oligonucleotides were prepared as previously described⁷. Crystals of form B were grown by hanging drop vapour diffusion at 16 °C using a protein:GGTTA molar ratio of 1:1.2, and precipitant/well

solution containing 100 mM trisodium citrate, pH 4.4, 16–18% polyethylene glycol 8,000 and 5 mM DTT. Crystals of form C were grown using a precipitant/well solution containing 100 mM Tris-HCl, pH 8.5, 27–30% polyethylene glycol 4,000 and 5 mM DTT with a protein:GGTTAC molar ratio of 1:1.5.

Data collection and structure determination

All crystals were flash-frozen in liquid nitrogen for storage and data collection under cryogenic conditions (100 K). Data were collected at the Advanced Light Source HHMI beamlines 8.2.1 and 8.2.2 using 1.0688 Å radiation and an ADSC Quantum Q4 detector, and processed using DENZO and SCALEPACK²⁵. SAD data from the gold derivative (form B) were used to obtain initial phases. Four gold atoms were located by anomalous Patterson and difference Fourier methods, using CNS²⁶. Gold positions were refined and the SAD phases calculated using CNS; the initial SAD map was improved by solvent flattening. A model was built into the density using O (ref. 27); the model was then transferred into the native unit cell by rigid-body refinement and further refined using simulated-annealing and positional refinement (CNS), with manual rebuilding (O). The structure of form C was determined by molecular replacement using the Pot1pN structure of form B as a partial search model (CNS). In the initial stage of the molecular replacement, the 'extra' electron density corresponding to the ssDNA (GGTTAC) was clearly defined. Model building and refinement were performed with O and CNS. Through the refinement, non-crystallographic symmetry restraints were applied to the six Pot1pN molecules in the asymmetric unit.

Equilibrium binding constants

Filter binding assays were performed as described previously⁷. Each oligonucleotide had seven A bases preceding the telomeric sequence to facilitate retention of non-protein-bound DNA to the second membrane, thereby improving quantification; this addition increases the *K*_D from 83 to 480 nM (ref. 7).

Analysis of Pot1 mutants

Point mutations were introduced by site-directed mutagenesis, verified by sequencing and subcloned into a 3.5-kilobase (kb) genomic fragment containing the Pot1 ORF and approximately 1 kb of flanking sequence on either side. Plasmids were introduced into a diploid (*pot1*^{+/pot1::kan}) *S. pombe* strain and selected for by omitting uracil from the growth medium. Diploids containing the plasmid were transferred to malt extract sporulation media and spores purified by glucusase treatment. Haploids were grown in the absence of uracil and replica-plated to media containing geneticin to test for the presence or absence of wild-type *pot1* at the genomic locus.

Received 25 June; accepted 22 September 2003; doi:10.1038/nature02092.

- Cervantes, R. B. & Lundblad, V. Mechanisms of chromosome-end protection. *Curr. Opin. Cell Biol.* **14**, 351–356 (2002).
- de Lange, T. Protection of mammalian telomeres. *Oncogene* **21**, 532–540 (2002).
- Baumann, P. & Cech, T. R. Pot1, the putative telomere end-binding protein in fission yeast and humans. *Science* **292**, 1171–1175 (2001).
- Baumann, P., Podell, E. & Cech, T. R. Human Pot1 (protection of telomeres) protein: cytoplasmic localization, gene structure, and alternative splicing. *Mol. Cell. Biol.* **22**, 8079–8087 (2002).
- Colgin, L. M., Baran, K., Baumann, P., Cech, T. R. & Reddel, R. R. Human POT1 is a positive regulator of telomere length. *Curr. Biol.* **13**, 942–946 (2003).
- Loayza, D. & de Lange, T. POT1 as a terminal transducer of TRF1 telomere length control. *Nature* **424**, 1013–1018 (2003).
- Lei, M., Baumann, P. & Cech, T. R. Cooperative binding of single-stranded telomeric DNA by the Pot1 protein of *Schizosaccharomyces pombe*. *Biochemistry* **41**, 14560–14568 (2002).
- Murzín, A. G. OB (oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J.* **12**, 861–867 (1993).
- Bryan, T. M. & Cech, T. R. Telomerase and the maintenance of chromosome ends. *Curr. Opin. Cell Biol.* **11**, 318–324 (1999).
- Gray, J. T., Celandier, D. W., Price, C. M. & Cech, T. R. Cloning and expression of genes for the *Oxytricha* telomere-binding protein: specific subunit interactions in the telomeric complex. *Cell* **67**, 807–814 (1991).
- Horvath, M. P., Schweiker, V. L., Bevilacqua, J. M., Ruggles, J. A. & Schultz, S. C. Crystal structure of the *Oxytricha nova* telomere end binding protein complexed with single strand DNA. *Cell* **95**, 963–974 (1998).
- Mitton-Fry, R. M., Anderson, E. M., Hughes, T. R., Lundblad, V. & Wuttke, D. S. Conserved structure for single-stranded telomeric DNA recognition. *Science* **296**, 145–147 (2002).
- Chandra, A., Hughes, T. R., Nugent, C. I. & Lundblad, V. Cdc13 both positively and negatively regulates telomere replication. *Genes Dev.* **15**, 404–414 (2001).
- Taggart, A. K., Teng, S. C. & Zakian, V. A. Est1p as a cell cycle-regulated activator of telomere-bound telomerase. *Science* **297**, 1023–1026 (2002).
- Theobald, D. L., Mitton-Fry, R. M. & Wuttke, D. S. Nucleic acid recognition by OB-fold proteins. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 115–133 (2003).
- Bochkarev, A., Pfuetzner, R. A., Edwards, A. M. & Frappier, L. Structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA. *Nature* **385**, 176–181 (1997).
- Raghuathan, S., Kozlov, A. G., Lohman, T. M. & Waksman, G. Structure of the DNA binding domain of *E. coli* SSB bound to ssDNA. *Nature Struct. Biol.* **7**, 648–652 (2000).
- Bastin-Shanower, S. A. & Brill, S. J. Functional analysis of the four DNA binding domains of replication protein A. The role of RPA2 in ssDNA binding. *J. Biol. Chem.* **276**, 36446–36453 (2001).
- Bochkarev, A., Korolev, S., Lees-Miller, S. P. & Bochkarev, A. Structure of the RPA trimerization core and its role in the multistep DNA-binding mechanism of RPA. *EMBO J.* **21**, 1855–1863 (2002).
- Hollis, T., Stattel, J. M., Walther, D. S., Richardson, C. C. & Ellenberger, T. Structure of the gene 2.5 protein, a single-stranded DNA binding protein encoded by bacteriophage T7. *Proc. Natl. Acad. Sci. USA* **98**, 9557–9562 (2001).

21. Classen, S., Lyons, D., Cech, T. R. & Schultz, S. C. Sequence-specific and 3'-end selective single-stranded DNA binding by *Oxytricha nova* telomere end binding protein (*OnTEBP*) α subunit. *Biochemistry* **42**, 9269–9277 (2003).
22. Peersen, O. B., Ruggles, J. A. & Schultz, S. C. Dimeric structure of the *Oxytricha nova* telomere end-binding protein α -subunit bound to ssDNA. *Nature Struct. Biol.* **9**, 182–187 (2002).
23. Classen, S., Ruggles, J. A. & Schultz, S. C. Crystal structure of the N-terminal domain of *Oxytricha nova* telomere end-binding protein α subunit both uncomplexed and complexed with telomeric ssDNA. *J. Mol. Biol.* **314**, 1113–1125 (2001).
24. Arcus, V. OB-fold domains: a snapshot of the evolution of sequence, structure and function. *Curr. Opin. Struct. Biol.* **12**, 794–801 (2002).
25. Otwinowski, Z. in *Proceedings of the CCP4 Study Weekend* (eds Sawyer, L., Isaacs, N. & Burley, S.) 56–62 (SERC Daresbury Laboratory, Daresbury, UK, 1993).
26. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
27. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these methods. *Acta Crystallogr. A* **47**, 110–119 (1991).
28. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structure. *J. Appl. Crystallogr.* **24**, 945–950 (1991).
29. Merritt, E. A. & Murphy, M. E. P. Raster 3D version 2.0, a program for photorealistic molecular

graphics. *Acta Crystallogr. D* **50**, 869–873 (1994).

30. Nicholls, A. & Honig, B. A rapid finite difference algorithm, utilizing successive overrelaxation to solve Poisson Boltzmann equation. *J. Comput. Chem.* **270**, 26184–26191 (1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank F. Guo, L. Chen and D. Theobald for insightful discussions, and G. Meigs and K. Henderson from ALS beamlines 8.2.1 and 8.2.2 for help in data collection. P.B. thanks W. Nudson and J. Bunch for technical assistance. This work was supported by the National Institutes of Health (T.R.C.) and by the Stowers Institute for Medical Research (P.B.). M.L. is a research fellow of the Helen Hay Whitney Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.R.C. (thomas.cech@colorado.edu). The atomic coordinates and structure factors have been deposited in the RCSB Protein Data Bank with accession codes 1QZG (form B) and 1QZH (form C).

Warming to the task

New for PCR, thermocyclers and DNA handling.

AccuPrime

Invitrogen www.invitrogen.com

Winning formula for high-fidelity PCR

Invitrogen's AccuPrime *Pfx* DNA polymerase is formulated to bring increased specificity to high-fidelity PCR. It combines 'Platinum' *Pfx* DNA polymerase, which has a 26-fold higher fidelity than *Taq*, with 'hot-start' antibodies and AccuPrime accessory proteins for accurate primer binding in every PCR cycle. The combination increases yield and reduces background.

Carbon fibre rotors for KR25i

Jouan www.jouan.com

Easy come, easy go

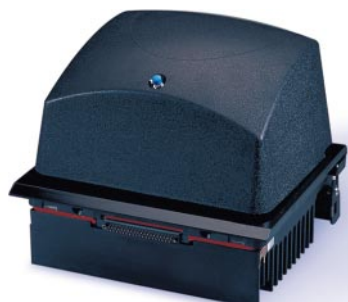
Jouan has a new range of lightweight carbon fibre rotors for its KR25i high-speed centrifuge. They can be coupled with the Auto-Lock rapid rotor exchange system for multiple protocols. The 6×500 ml (11,000 r.p.m., 21,374g) and 6×250 ml (14,000 r.p.m., 30,074g) rotors are up to 50% lighter than comparable aluminium rotors, making them easier to handle, and can be exchanged in less than 10 seconds. Cleaning and disinfecting are simplified by the use of carbon fibre, which is not subject to the same corrosion risk as aluminium. The rotors are especially suitable for differential centrifugation for the harvesting of subcellular particles as they make it easier to accommodate samples of different volumes for different protocols. Samples of up to 500 ml can be processed using a six-place rotor for high-volume protocols. For lower volumes the rotor can be removed easily and replaced with one to process 50 ml samples at 63,000g.

Moto Alpha sample block

MJ Research www.mjresearch.com

Set your seal on it

The Moto Alpha interchangeable sample block is designed for use with any thermal cycler in the 'DNA Engine' line. It provides



Block booking: for DNA Engines.

reliable plate positioning and sealing and can be integrated with a robotic system or used as a stand-alone unit. Sealing force is controlled by the user through a software program and is actively monitored for reproducible, high-pressure sealing. Evaporative loss is minimized, even for low-volume cycling reactions. A non-stick block with a plate lifter and lid-mounted ejectors ensure that the plate is always positioned for easy removal.

Human Universal Reference Total cDNA

BD Biosciences Clontech www.clontech.com

Broader coverage

Human Universal Reference Total cDNA can be used to compare data from different quantitative PCR experiments. As it is prepared from a total RNA pool collected from several different human tissues instead of cell lines, it provides broader gene coverage. The company says that its total RNA is virtually free of genomic DNA, allowing for accurate measurement of transcript copy number. Both high- and low-abundance genes are well represented, enabling researchers to prepare a wide range of serially diluted standards for each assay.

HS400 compact hybridization station

Tecan www.tecan.com

For modest researchers

Designed for those who need to carry out a modest number of experiments per run, the fully automated HS400 hybridization station from Tecan is said to improve the reproducibility of microarray results. It enables fully automated hybridization of up to four microarray slides, which can be loaded simultaneously. Incubations may be performed in the range 4–85°C. The unit offers efficient liquid agitation and on-board slide drying to maximize the quality of the results and shorten hybridization times. It can be operated via the integrated touch-screen module or from an external PC. Applications include gene expression profiling, gene regulation, pharmacogenomic studies, identification of new drug targets and toxicology.

Silencer Express

Ambion www.ambion.com

Kits with SECs appeal

These kits from Ambion provide a rapid, PCR-based method of preparing siRNA expression cassettes (SECs) that can be introduced into cells to induce gene silencing without prior cloning or sequencing. SECs are easy to prepare and are thus especially suitable for

screening siRNA sequences and identifying the most effective promoter in an experimental system. Once an effective siRNA sequence has been identified, the SEC can be sub-cloned into an appropriate vector for long-term gene silencing studies.

Phusion DNA polymerase

MJ Research www.mjr.com

High fidelity at any length

Phusion high-fidelity DNA polymerase incorporates a nonspecific DNA binding domain that lends physical stability to the polymerase–DNA complex, enabling replication of long templates with improved accuracy, speed and robustness. The product is said to optimize fidelity, reliability and yield and is especially useful with long or difficult templates.

RapidPURE

Q-Biogene www.qbiogene.com

DNA at the ready

Q-Biogene has introduced an easy-to-use kit for rapid alkaline lysis and purification of plasmid DNA from *E. coli*. The BIO 101 Systems RapidPURE Plasmid Mini Kit provides all the tools and reagents necessary to quickly prepare plasmid DNA for transfection, sequencing and amplification. Efficiency and yield are enhanced by the 'Turbo' filter, which binds up to 20 µg of plasmid DNA and is more convenient to use than silica slurries or spin columns, and by 'CIRCLEGROW', a high-yield growth media that provides cell growth and plasmid replication well beyond log phase. The RapidPURE kit is available in 60, 120 and 300 preparation sizes.

7-deaza-8-aza-dA amidites

Cambio www.cambio.co.uk

Expanded potential for G-rich sequencing

G-rich sequences in DNA can make PCR more difficult, sequencing may present problems and DNA structures are disrupted by inter- and intra-strand hydrogen bonding. Probes made from G-rich sequences can be rendered inaccessible by hydrogen bonds at the N7 position of guanine. A novel solution is to move the 7 nitrogen to the 8 position. This new amidite, together with the equivalent 7-deaza-8-aza-dA, raises the T_m above the standard A–T base pair. It is described by the manufacturer as a very useful base for researchers wishing to modify G-rich and A-rich sequences to make better primers and probes in areas previously considered inaccessible.

HotMaster

Eppendorf www.eppendorf.com

Hot start meets cold stop

Most 'hot-start' systems either block the DNA polymerase by antibodies or prompt

deactivation of the enzyme using a chemical modification, and require an activation time (denaturing step) of 5–15 minutes in order to transfer to active status. Eppendorf's HotMaster kit uses an inert ligand that disconnects immediately from the active centre of the polymerase at higher temperatures. The activation step is superfluous. The system also contains a 'cold stop' feature, in which the ligand 'docks' onto the polymerase again, deactivating it when the temperature descends below a critical threshold. This feature increases the specificity of the PCR reaction.



HotMaster: hot start PCR with a twist.

MCA System

Thermo Hybaid www.thermo.com

Melting curves have many uses

Thermo Hybaid's MCA (melting curve analysis) system is based on the technology used in its DASH instrument for genotyping. The MCA system features embedded and front-end software that will enable researchers to design a protocol within the temperature range of 20–100°C. Up to eight excitation and eight emission filters per 96-well plate reading can be chosen, allowing a number of different fluorescent systems to be used. The software logs all data from the system and produces differential melting curve peaks that are then autoscored and designated. MCA is proving to be the method of choice for a range of techniques, including mutation detection, post-PCR analysis, DNA methylation profiling, SNP scoring and detection by heteroduplex analysis. This system is designed with the flexibility to cater for all of these techniques.

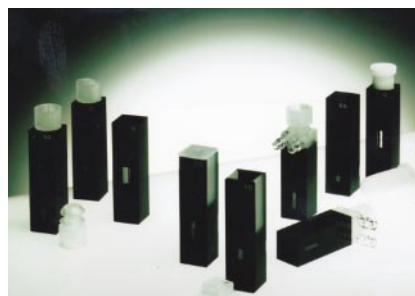
CST DNA extraction kit

DNA Research Innovations www.dri-dna.com

Switch on to new technology

This easily automated DNA purification kit for extracting genomic DNA from blood is designed for use with EDTA or ACD treated blood and standard buffer coat preparations. It uses charge switch technology (CST), a chemical process that enables researchers to purify nucleic acids without the use of harsh ionic chaotropes, hazardous organic solvents and expensive alcohols, all of which have been shown to inhibit downstream processes such as PCR. CST works as a pH-dependent ionic

switch that selectively binds nucleic acids when 'switched on' in the presence of optimized reagents. When 'switched off' after washing, the charge is neutralized, allowing purified DNA to be eluted. This technology allows high-quality DNA to be purified easily. The kit can be readily automated on most liquid-handling robots since the protocols require no centrifugation or vacuum filtration. Applications include mutation analysis, medical diagnosis, studies of population dynamics and forensic typing.



Downsize: Starna Type 26 extra-small cuvettes.

Starna Type 26 LHS sub-micro cells

Optiglass www.optiglass.co.uk

These leave no room for error

DNA denaturation studies require samples to be subjected to temperatures of around 80°C while the change in absorbance is measured using an ultraviolet spectrophotometer. At these temperatures, particularly with very small samples, evaporation can cause significant measurement errors due to changes in volume and, hence, concentration. In a normal cuvette with limited sample, an empty head-space will exist above the liquid surface into which evaporation will occur. The most commonly used, but cumbersome, solution is to cover the sample with a high boiling point, water-immiscible solvent to act as an evaporation barrier. Starna Type 26 LHS sub-micro cells minimize evaporation losses by more than 95% compared with conventional designs. This is achieved by eliminating the head-space above the sample using a specially

profiled Teflon stopper. Cells are available with nominal sample volumes from 10 µl to 440 µl, with fully heat-fused Suprasil quartz windows as standard. The black, self-masking cell body can be supplied to accommodate different instrument optical beam heights to maximize energy transmission with the smallest cell apertures. The company also supplies a Type 64 (160 µl) constant temperature cell that allows for fluid circulation around the sample chamber.

PCR tools

SuperArray Bioscience www.superarray.com

Two new ideas

Making PCR easier is the aim for these new products from SuperArray. The Reaction-Ready Sweet PCR master mix is a lyophilized cocktail containing an orange tracking dye allowing the user to transfer reactions directly from PCR tubes to agarose gel wells. All other PCR components are also included in the mix. Just add water, template and primers, say the manufacturers. One kit contains enough PCR master mix for 98 reactions. The second new introduction is a range of SingleGene PCR Kits designed to make it convenient to examine the expression of individual genes by RT-PCR. Each kit contains the Sweet PCR master mix and a pre-tested, validated gene-specific primer set that yields a single band of the predicted size. There is no need to test several different custom primer pairs, and enough reagents are included for 24 reactions.

NucleasEliminator Spray

AMRESKO www.amresco-inc.com

Something in the air

AMRESKO's NucleasEliminator decontamination liquid now comes in an easy-to-use spray bottle. The new NucleasEliminator Spray makes it possible for anyone working with RNA to quickly cover and clean any large DNase/RNase contaminated lab surface (benches, beakers and so on) with minimal effort. Just aim and pull the trigger. The new spray has the same nuclease removal properties as previous incarnations of NucleasEliminator.

EchoTherm

Torrey Pines Scientific

www.torrey-pinescientific.com

Bath time

These new EchoTherm Models, the SC20 Digital and SC25 Fully Programmable Dry Baths, are orbital mixers providing chilling and heating for biological and other samples. The SC20 is a simple digital unit and the SC25 is a fully programmable unit with 5-program memory capacity. Both units have a temperature range from -10°C to 100°C and incorporate a variable-speed orbital mixer which



Home and dry: EchoTherm for PCR.

allows for mixing and controlling temperature of samples simultaneously. Both units have 30-day countdown timers with alarm and auto-off. The SC20 and SC25 accommodate accessory sample blocks available for 0.2, 0.5, 1.5 and 15 ml centrifuge tubes, 2 ml vials, 20 ml scintillation vials, PCR tubes and plates, 96-well and 384-well assay plates of all shapes and other blocks for various sizes of test tubes. The units are Peltier driven, control to 1°C, mix from 200 to 1,000 r.p.m. and have a backlit two-line alphanumeric display. Both models can be used to run temperature/time profiles, unattended restriction digestions or ligations, automatic enzyme reactions and deactivations, storing oocytes at 17°C, storing DNA libraries at the workstation, and for replacing messy ice buckets and more.

GeneAmp AccuRT kits

Applied Biosystems

www.appliedbiosystems.com

Raising the temperature

A new designer DNA polymerase from Applied Biosystems makes it possible to perform Mg²⁺-activated, single-enzyme RT-PCR in a one-step, single-tube-reaction. The GeneAmp AccuRT RNA PCR enzyme is thermostable and thermoactive, allowing reverse transcription and PCR amplification to be performed at elevated temperatures. This helps to remove secondary structures from the template and facilitates highly stringent, accurate annealing of gene-specific primers. For hot start reactions, there is also a new GeneAmp AccuRT Hot Start RNA PCR Kit. An aptamer-based system, it provides hot start capability for both reverse transcription and PCR. This hot start process is reversible, providing a 'cold stop' when the reaction is complete. The GeneAmp AccuRT RNA PCR enzyme, the GeneAmp AccuRT Hot Start RNA PCR enzyme and their associated buffers can also be purchased separately.

These notes are compiled in the Nature office from information provided by the manufacturers. Material intended for publication can be sent to newproducts@nature.com.

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)

Netherlands/ Italy/ Spain/ Portugal/ Belgium:

Evelina Rubio Hakansson (4973)
Scandinavia: Sille Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Natureevents:

Paul Constant (4954)

Production Manager:

Billie Franklin
To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@nature.com

US Sales Manager:

Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harakakamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@nature.com

naturejobs

A question of gender

The statistics agree that women are disproportionately represented in science (see *Nature* 426, 8; 2003). But there is little agreement on how to redress the balance. Some plans call for legislation, others for quotas. Some advocates hope that female pioneers will lead the way, while others worry that high-profile successes will be accused of getting breaks because of their gender.

But the first step is to recognize a problem. The European Commission (EC) has taken that step and is trying to address it by asking grant applicants to provide gender-equality plans for their research teams in the current round of EC funding. Many scientists, however, are uncomfortable with the idea of legislating for gender equality (see page 210). Similarly, when European research centres have tried to target women for their ranks (see page 211), their projects attracted many qualified applicants — but also charges of reverse discrimination, inevitable in any quota system.

Even the pioneers disagree. Persis Drell became director of research at the Stanford Linear Accelerator Center in California last year, but says that she has reservations about being regarded as a role model for her gender. Biologist Shirley Tilghman, president of Princeton University, New Jersey, since 2001, takes the opposite approach, canvassing for more women in science.

As women scientists in Europe begin to rise to the top of their careers, they must grapple with similar issues. But at least that is starting to happen. Harriet Wallberg-Henriksson will become the first female president of Stockholm's Karolinska Institute next year.

Yet ultimately, it may be that market forces resolve the problem. If, as planned, Europe does increase its investment in research and development to 3% of gross domestic product by 2010, there won't be enough male scientists to do the work.

Paul Smaglik
Naturejobs editor



Contents

SPECIAL REPORT

Europe attempts to promote
women scientists p210

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Europe is pushing to get more women scientists into industry and academia, but can the commission legislate for gender equality? Sally Goodman investigates.

Choose your husband carefully: one willing to share the laundry and make the children's supper. That was one piece of advice for women researchers at a recent conference in Berlin on women in industrial research. It's an example of how gender equality needs to be integrated into every part of public and private life if more women are to reach their potential.

The reasons for the shortage of women in science and engineering, particularly in senior positions, are well documented: the 'leaky pipeline' caused by opaque recruitment practices, the pressures of juggling workload with family commitments, systems of promotion that work against women who interrupt their careers to have children, to name but a few.

Responses to these problems have often treated the symptoms rather than the causes, through schemes to push more women into the pipeline. But the European Commission (EC) is now spearheading efforts to promote deep-seated organizational and cultural change in both private- and public-sector research.

The EC is pushing for industry to hire and promote more female scientists and is using the Sixth Framework Programme to make similar inroads in academia. But for real progress, individual nations, universities and companies will need to rethink their own practices.

At a Berlin conference on 10–11 October, jointly organized by the EC and the German education and research ministry, industry and governments were challenged to speed up the necessary changes.

Little information existed on the subject before the publication in January of an EC-commissioned report by a group of industrial and academic experts: *Women in Industrial Research: A Wake-up Call for European Industry*. This revealed that fewer than 15% of the industrial researchers in Europe are women — and only 9.6% in Germany, where a third of Europe's industrial researchers are based.

Meeting Europe's target of investing 3% of gross domestic product in research and development by 2010 will mean recruiting up to 700,000 more researchers,



Rosalind Franklin: 45 years after her death, women are still struggling for recognition.

mainly in the private sector — and highly qualified women are a rich, untapped resource.

Several companies have woken up to this reality. In Berlin, a group of seven Europe-based companies announced a public commitment to improving the lot of women in industrial R&D. Airbus, Air Liquide, EADS, Hewlett-Packard, Rolls-Royce, Schlumberger and Siemens have each pledged to take measures such as monitoring targets and sponsoring role models, for example by endowing university chairs. "We hope other companies will take up the challenge and join us," says Andrew Gould, chief executive of energy company Schlumberger, who led the initiative.

"The business case for gender diversity is obvious and essential." Others agree. "Diversity breeds creativity and innovation — factors that are a vital part of research," says Jenny Holmes, R&D diversity director at drug company AstraZeneca.

Also, women in product development are more likely to take female consumers' needs into account — a point made by cosmetics firm L'Oréal, which boasts 55% of women on its research staff. L'Oréal also organizes annual 'For Women in Science' awards and fellowships in collaboration with UNESCO.

FRAMEWORK FIX

Although the percentage of women researchers in Europe's public sector is double that in industry, the situation is still far from satisfactory. The EC hopes to set a good example by integrating equality into the latest Framework programme for research. Funding proposals for integrated projects and networks must, for the first time, include an action plan to promote gender equality — tackling both research content and the promotion of women's participation in research.

But an informal survey by the EC's Women and Science Unit of the first-round applicants for funding illustrates that much work remains to be done. Only 14% of lead scientists were women. And only 10% of projects accepted for funding detailed their gender

action plan, with 15% not addressing the issue at all.

Petra Bender is diversity manager at the Research Centre Jülich in Germany, a multidisciplinary public-research centre that runs positive-action programmes for women researchers. She says that despite her efforts to explain how gender impact could be written into funding proposals, some of her colleagues remain perplexed. "Many researchers just don't understand the concept of mainstreaming," she says, adding that a typical comment was: "But crystals don't have gender."

Those projects that did make an effort to develop a gender action plan included target-setting to achieve gender balance in project management positions, plans to develop mentoring schemes for younger researchers and organizing outreach activities for schools. Many of these could easily be adopted by future applicants in need of inspiration. Not surprisingly, the promotion of women in research was better addressed than the gender implications of research which, for many projects outside the life sciences, is not immediately obvious.

Framework projects selected for funding will now be encouraged to improve their gender action plans as

part of contract negotiations, and the best ideas will be shared. "We're not looking for perfection in the proposals. It's a process that will take time, but at least the visibility of the issue has been achieved," says Nicole Dewandre, head of the Women and Science Unit.

NATIONAL GENDER IDENTITY

But with the EC funding only 5% of civil research in Europe, national governments and companies must take up the gauntlet if rapid change in women's participation in science is to happen. Above all, they need to be accountable. As one speaker at the Berlin conference put it: "What gets measured, gets done."

Putting more women into senior positions will not be enough on its own. "We need to move beyond numbers to a cultural change in organizations that is truly diverse, inclusive and accommodating," says Teresa Rees, from the school of social sciences at the University of Cardiff, UK, and rapporteur of the EC's expert group on women in industrial research.

Sally Goodman is a freelance science writer based in Paris.

The quota conundrum

Sometimes you have to break the law to get things right. Germany's Research Centre Jülich pushed the boundaries with positive discrimination in 1999. Now the University of Groningen in the Netherlands has intentionally flouted anti-quota laws.

Last year, Groningen's natural-sciences faculty set up the Rosalind Franklin Fellowships, offering five women-only tenure-track positions — even though positive discrimination is illegal in the Netherlands.

Despite the legal transgression no one protested, says mathematician Ruth Curtain, chair of the search committee: "The academic world in the Netherlands is very embarrassed about gender discrimination."

Requirements for the Franklin fellowships — named after the British X-ray crystallographer whose work contributed to the discovery of DNA structure — included postdoctoral experience abroad, publications in top journals and proof of international recognition.

The scheme drew 112 applications from around the world and the first fellows started work this autumn. After five years, if their evaluations are successful, they will be offered permanent professorships.

"We want to send a clear message that there are opportunities for talented women, beyond a string of temporary, postdoc appointments," says Curtain. Women were attracted by the fellowships' broad and nonspecific range of research fields, she says.

"When you advertise positions for a specific field, you often have only a few women applying, because in many fields the imbalance starts from the lower levels," Curtain says, adding that it is also important to let women set up their own lines of research.

The dean of the faculty, Douwe Wiersma, supported the programme, though many of his male colleagues disapproved. "The majority believe there is no discrimination in selection committees, and that gender inequality will redress itself naturally as more women enter the science field," Wiersma says.

In reality, says Curtain, qualified women don't get short-listed for higher positions — despite the rising number of women PhDs during the past two decades.

"The usual answer is: 'We would like to appoint women, but they simply do not apply,'" says Curtain. The response to the Rosalind Franklin programme disproves that notion, she points out.

In Germany, the Research Centre Jülich was also criticized by some when it advertised tenure-track positions for women (see *Nature* 398, 550; 1999).

"We offered only three positions a year so of course you can't change the statistics with that alone," says Petra Bender, the centre's diversity manager. But the scheme had a knock-on effect, with numbers of women researchers at the institute doubling to 15% since 1998. "It has led to more women applying for other jobs because they now see us as an employer who may offer opportunities to young women," she says.

Subsequently, the University of Munich proposed a scheme to increase universities' funding if they put more women in high positions. The Bavarian research ministry refused, saying that female presence in science must develop "naturally".

Nicola Nosengo recently completed an internship in *Nature's* Munich office.

MICHEL DE GROOT



Franklin fellows, from left: Charlotte Hemelrijk, Heide Gluesing-Leurssen, Petra van Koningsbruggen, Elisabetta Pallante and Beatriz Noheda.